Technical Report 776

Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS

Jody L. Toquam, Jeffrey J. McHenry, VyVy A. Corpe, Sharon R. Rose, Steven E. Lammlein, Edward Kemery, Walter C. Borman, Raymond Mendel, and Michael J. Bosshardt

Personnel Decisions Research Institute

Selection and Classification Technical Area

Manpower and Personnel Research Laboratory

U. S. Army

Research Institute for the Behavioral and Social Sciences

January 1988

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
COL, IN
Commanding

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Jane Arabian
Paul Rossmeissl

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) --	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 776
6a. NAME OF PERFORMING ORGANIZATION Human Resources Research Organization 6b. OFFICE SYMBOL (If applicable) HumRRO	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS (City, State, and ZIP Code) 1100 South Washington Street Alexandria, Virginia 22314-4499	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, Virginia 22333-5600
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA 903-82-C-0531
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. 2Q263-731A792 / TASK NO. 2.3.2 / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS

12. PERSONAL AUTHOR(S) Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S.E., Kemery, E., Borman, W. C., Mendel, R., and Bosshardt, M. J. (PDRI)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM Oct 1983 TO Sep 1985	January 1988	109

16. SUPPLEMENTARY NOTATION

Lawrence M. Hanser, Contracting Officer's Representative.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Behavioral scales, Classification, Criterion measures, First-term evaluation, MOS-specific tests, Performance dimensions, Performance ratings, Project A Field Test, (continued)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The research described in this report was performed under Project A, the U.S. Army's current, large-scale, manpower and personnel effort to improve the selection, classification, and utilization of Army enlisted personnel. This report documents the development and field test of behaviorally anchored rating scales for nine Military Occupational Specialties (MOS). These include combat, combat support, and noncombat MOS.

For each MOS, the behavioral analysis method was used to generate examples of performance. These examples were used to identify performance effectiveness dimensions and to develop behavioral definitions of performance for each dimension. Across the nine MOS, behavioral summary rating scales contained from 7 to 13 performance dimensions.

The nine sets of MOS-specific behavioral summary rating scales were field tested in continental United States and overseas locations in two groupings (Batch A and (continued)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Lawrence M. Hanser	22b. TELEPHONE (Include Area Code) (202) 274-8275 / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

ARI Technical Report 776

18. Subject Terms (continued)

Selection, Soldier effectiveness

19. Abstract (continued)

Batch B). For each MOS, ratings scales were administered to 120 to 160 first-term soldiers and their supervisors.

Within each MOS, interrater reliability estimates for individual performance dimension ratings were reasonably high and rating distributions were acceptable, indicating no leniency or severity effects. Results from the field tests, along with suggestions from proponent review committees and Project A staff, were used to modify and prepare the nine sets of rating scales for the Concurrent Validation study.

The appendixes that provide further documentation for this research consist of the materials developed for each of the nine MOS. They are issued in a separate report, with limited distribution, as follows:

ARI Research Note, Appendixes to ARI Technical Report: Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS (in preparation).

Volume 1 - Appendix A, MOS 13B
Appendix B, MOS 64C

Volume 2 - Appendix C, MOS 71L
Appendix D, MOS 95B

Volume 3 - Appendix E, MOS 11B
Appendix F, MOS 19E
Appendix G, MOS 31C

Volume 4 - Appendix H, MOS 63B
Appendix I, MOS 91A

Technical Report 776

Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS

Jody L. Toquam, Jeffrey J. McHenry, VyVy A. Corpe, Sharon R. Rose, Steven E. Lammlein, Edward Kemery, Walter C. Borman, Raymond Mendel, and Michael J. Bosshardt
Personnel Decisions Research Institute

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

January 1988

Army Project Number
2Q263731A792

Manpower and Personnel

FOREWORD

This document describes the development and field testing of behaviorally anchored rating scales for evaluating performance of first-term personnel in nine Military Occupational Specialties (MOS). The research was part of Project A, the Army's current, large-scale manpower and personnel effort to improve the selection, classification, and utilization of Army enlisted personnel. The thrust for the project came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training and performance.

Project A is being conducted under contract to the Selection and Classification Technical Area (SCTA) of the Manpower and Personnel Research Laboratory (MPRL) at the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The portion of the effort described herein is devoted to the development and validation of Army Selection and Classification Measures, and referred to as "Project A." This research supports the MPRL and SCTA mission to improve the Army's capability to select and classify its applicants for enlistment or reenlistment by ensuring that fair and valid measures are developed to evaluate applicant potential based on expected job performance and utility to the Army.

Project A was authorized through a Letter, DCSOPS, "Army Research Project to Validate the Predictive Value of the Armed Services Vocational Aptitude Battery," effective 19 November 1980; and a Memorandum, Assistant Secretary of Defense (MRA&L), "Enlistment Standards," effective 11 September 1980.

In order to ensure that Project A research achieves its full scientific potential and will be maximally useful to the Army, a governance advisory group comprised of Army general officers; interservice scientists; and experts in personnel measurement, selection, and classification was established. Members of the latter component provide guidance on technical aspects of the research, while general officer and interservice components oversee the entire research effort; provide military judgment; periodically review research progress, results, and plans; and coordinate within their commands. Members of the General Officer's Advisory Group include MG Porter (DMPM) (Chair), MG Briggs (FORSCOM, DCSPER), MG Knudson (DCSOPS), BG Franks (USAREUR, ADCSOPS), and MG Edmonds (TRADOC, DCS-T). The General Officer's Advisory Group was briefed in May 1985 on the issue of obtaining proponent concurrence of the criterion measures before administering the concurrent validation. Members of Project A's Scientific Advisory Group (SAG), who guide the technical quality of the research, include Drs. Milton Hakel (Chair), Philip Bobko, Thomas Cook, Lloyd Humphreys, Robert Linn, Mary Tenopyr, and Jay Uhlaner. The SAG was briefed in October 1984 on the results of the Batch A field test administration. Further, the SAG was briefed in March 1985 on the contents of the proposed Trial Battery.

A comprehensive set of new selection/classification tests and job performance/training criteria have been developed and field tested. Results from the Project A field tests and subsequent concurrent validation will be used

to link enlistment standards to required job performance standards and to more accurately assign soldiers to Army jobs.

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

Authors contributing to this report participated by writing chapters and/or developing behavioral summary scales for one or more Military Occupational Specialties (MOS). The authors extend their thanks to the many Project A staff who assisted in developing, field testing, and modifying the MOS-specific behavioral summary rating scales.

Glenn Hallum, Cynthia Owens-Kurtz, Mary Ann Hanson, Cheryl Paullin, and Teresa Russell of Personnel Decisions Research Institute (PDRI) assisted in all phases of rating scale development. James Harris of the Human Resources Research Organization (HumRRO) scheduled the workshops and planned and prepared the field test data collection trips. Dr. Lauress Wise and Winnie Young of the American Institutes for Research (AIR) compiled and analyzed the data reported in this document. Dr. Michael Rumsey of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) presented the rating scales to the Proponent Review committees. In addition, numerous Project A staff involved in Task 5 from ARI, AIR, HumRRO, and PDRI provided comments and suggestions for modifying and improving the MOS-specific rating scales.

Finally and most especially, we thank the many soldiers who contributed to this study. The Army points-of-contact (POC) at each post provided enormous assistance by arranging and scheduling workshops and field test data collection sessions. Perhaps the most important contributors were the first-term soldiers and their supervisors who participated in the behavioral analysis workshops and field test sessions. The conscientious efforts of all of these individuals are greatly appreciated.

DEVELOPMENT AND FIELD TEST OF BEHAVIORALLY ANCHORED RATING SCALES FOR NINE MOS

EXECUTIVE SUMMARY

Requirement:

Project A is a large-scale, multiyear research program intended to improve the selection and classification system for initial assignment of persons to U.S. Army Military Occupational Specialties (MOS). Specifically, Project A is to validate new and existing selection measures against both existing and project-developed criteria.

This report describes the development and field test of behaviorally anchored rating scales designed for nine MOS. These include infantryman (11B), Cannon Crewman (13B), Armor Crewman (19E), Single-Channel Radio Operators (31C), Light-Wheel Vehicle Mechanics (63B), Motor Transport Operators (64C), Administrative Specialists (71L), Medical Specialists (91A), and Military Police (95B).

Procedure:

For each MOS, the behavioral analysis method was used to generate examples of effective, average, and ineffective job performance. These examples were used to identify performance effectiveness dimensions and to develop behavioral definitions and standard of performance for each dimension. Across the nine MOS, behavioral summary rating scales contained from 7 to 13 performance dimensions.

These rating scales were field tested in continental United States and overseas locations. The first (Batch A) field test focused on four MOS, and the second (Batch B) field test focused on five MOS. For each MOS, rating scales were administered to 120 to 160 first-term soldiers and their supervisors.

Findings:

Results of the field test were encouraging. In particular, rating session administrators reported that participants understood and complied with instructions and found the rating scales useful for evaluating job performance; interrater reliability estimates were reasonably high; and rating distributions were acceptable with mean values slightly above the midpoint.

Utilization of Findings:

The MOS-specific rating scales will be administered in the Project A Concurrent Validation study scheduled for Summer 1985. Scores from these scales along with other scores from other criterion measures will be used to assess

the validity of existing and new selection measures. Information obtained from the field tests was used to modify, refine, and prepare the MOS-specific rating scales for the Concurrent Validity study. Overall, the scales required very few changes.

DEVELOPMENT AND FIELD TEST OF BEHAVIORALLY ANCHORED RATING SCALES FOR NINE MOS

CONTENTS

CONTENTS (Continued)

LIST OF APPENDIXES*

*The Appendixes are issued in a separate report, with limited distribution: ARI Research Note, Appendixes to ARI Technical Report: Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS (in preparation). Volume 1 contains the materials for Cannon Crewman (MOS 13B) and Motor Transport Operator (64C); Volume 2, Administrative Specialist (71L) and Military Police (95B); Volume 3, Infantryman (11B), Armor Crewman (19E), and Radio Teletype Operator (31C); Volume 4, Light-Wheel Vehicle Mechanic (63B) and Medical Specialist (91A).

Page

LIST OF TABLES

CONTENTS (Continued)

LIST OF FIGURES

OVERVIEW OF PROJECT A

Project A is a comprehensive long-range research and development program which the U.S. Army has undertaken to develop an improved personnel selection and classification system for enlisted personnel. The Army's goal is to increase its effectiveness in matching first-tour enlisted manpower requirements with available personnel resources, through use of new and improved selection/classification tests which will validly predict carefully developed measures of job performance. The project addresses the 675,000-person enlisted personnel system of the Army, encompassing several hundred different military occupations.

This research program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research effort that would be needed to develop the desired system. In 1982 a consortium led by the Human Resources Research Organization (HumRRO) and including the American Institutes for Research (AIR) and the Personnel Decisions Research Institute (PDRI) was selected by ARI to undertake the 9-year project. The total project utilizes the services of 40 to 50 ARI and consortium researchers working collegially in a variety of specialties, such as industrial and organizational psychology, operations research, management science, and computer science.

The specific objectives of Project A are to:

- Validate existing selection measures against both existing and project-developed criteria. The latter are to include both Army-wide job performance measures based on newly developed rating scales, and direct hands-on measures of MOS-specific task performance.
- Develop and validate new selection and classification measures.
- Validate intermediate criteria, such as performance in training, as predictors of later criteria, such as job performance ratings, so that better informed reassignment and promotion decisions can be made throughout a soldier's career.
- Determine the relative utility to the Army of different performance levels across MOS.
- Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

The research design for the project incorporates three main stages of data collection and analysis in an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria). In the first iteration, file data from Army accessions in fiscal years (FY) 1981 and 1982 were evaluated to explore the relationships between the scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB), and

their subsequent performance in training and their scores on the first-tour Skill Qualification Tests (SQT).

In the second iteration, a concurrent validation design will be executed with FY85 accessions. As part of the preparation for the Concurrent Validation, a "preliminary battery" of perceptual, spatial, temperament/personality, interest, and biodata predictor measures was assembled and used to test several thousand soldiers as they entered in four Military Occupational Specialties (MOS) in FY83/84. The data from this "preliminary battery sample" along with information from a large-scale literature review and a set of structured, expert judgments were then used to identify "best bet" measures. These "best bet" measures were developed, pilot tested, and refined. The refined test battery was then field tested to assess reliabilities, "fakability," practice effects, and so forth. The resulting predictor battery, now called the "Trial Battery," which includes computer-administered perceptual and psychomotor measures, is being administered together with a comprehensive set of job performance indices based on job knowledge tests, hands-on job samples, and performance rating measures in the Concurrent Validation.

Based partly on the results of the Concurrent Validation, the "Trial Battery" will be revised to become the "Experimental Predictor Battery" which in turn will be administered as part of the longitudinal validation stage beginning in the late Summer and early Fall of 1986.

For both the concurrent and longitudinal validations, a sample of 19 MOS were specially selected as representative of the Army's 250+ entry-level MOS. The selection was based on an initial clustering of MOS derived from rated similarities of job content. These 19 MOS account for about 45 percent of Army accessions. Sample sizes are sufficient so that race and sex fairness can be empirically evaluated in most MOS.

In the third iteration (the longitudinal validation), all of the measures, refined on the basis of experience in field testing and the Concurrent Validation, will be administered in a true predictive validity design. About 50,000 soldiers across 20 MOS will be included in the FY86-87 "Experimental Predictor Battery" administration and subsequent first-tour measurement. About 3500 of these soldiers are estimated for availability for second-tour performance measurement in FY91.

Activities and progress during the first two years of the project were reported for FY83 in ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37, and for FY84 in ARI Research Report 1393 and its related reports, ARI Technical Report 660 and ARI Research Note 85-14. Other publications on specific activities during those years are listed in those annual reports. The annual report on project-wide activities during FY85 is under preparation.

For administrative purposes, Project A is divided into five research tasks:

Task 1 -- Validity Analyses and Data Base Management
Task 2 -- Developing Predictors of job Performance
Task 3 -- Developing Measures of School/Training Success
Task 4 -- Developing Measures of Army-Wide Performance
Task 5 -- Developing MOS-Specific Performance Measures

The development and revision of the wide variety of predictor and criterion measures reached the stage of extensive field testing during FY84 and the first half of FY85. These field tests resulted in the formulation of the test batteries that will be used in the comprehensive Concurrent Validation program which is being initiated in FY85.

The present report is one of five reports prepared under Tasks 2-5 to report the development of the measures and the results of the field tests, and to describe the measures to be used in Concurrent Validation. The five reports are:

Task 2 -- "Development and Field Test of the Trial battery for Project A," Norman G. Peterson, Editor, ARI Technical Report 739, May 1987.

Task 3 -- "Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS," Robert H. Davis et al., ARI Technical Report in preparation.

Task 4 -- "Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program," Elaine D. Pulakos and Walter C. Borman, Editors, ARI Technical Report 716, July 1986.

Task 5 -- "Development and Field Test of Task-Based MOS-Specific Criterion Measures," Charlotte H. Campbell et al., ARI Technical Report 717, July 1986.

-- "Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS," Jody L. Toquam et al., ARI Technical Report in preparation.

CHAPTER 1: DEVELOPMENT OF BEHAVIORALLY ANCHORED RATING SCALES (BARS)

Objective

The U.S. Army is examining the effectiveness of its selection and classification battery, the Armed Services Vocational Aptitude Battery, in predicting training and job performance outcomes. As part of Project A, new predictor measures have been developed to supplement the current military selection and classification battery. Thus, an important feature of this project involves developing measures of training outcomes and job performance that can be used to estimate the validity of the ASVAB and the incremental validities of the new measures. The first wave of research activities has focused on first-term enlistee training and job performance outcomes.

Components of first-term enlistee job performance include measures of Army-wide, or general soldier effectiveness and measures of occupation-specific job requirements. These latter measures are the focus of Task 5 of Project A and of this report.

There are several ways to define the performance domain and to assess performance in MOS-specific job areas. For example, performance may be defined by the major or critical tasks comprising the job. Performance on such tasks may be assessed by measures that simulate critical activities of the job (e.g., hands-on tests), written tests that measure incumbents' knowledge of the critical components of the job (e.g., job knowledge tests), or measures that ask persons familiar with target incumbents to evaluate incumbents' performance in the task areas, using specially designed rating scales.

Another means of assessing performance involves identifying broad dimensions that define the critical job performance requirements. These dimensions may then be used to develop rating scales that measure performance effectiveness more broadly than task-oriented assessment instruments. Once again persons familiar with target incumbents are asked to evaluate incumbents' performance, using these rating scales.

For Task 5, both approaches have been used to measure job performance. That is, instruments assessing performance or knowledge in critical task areas and assessing performance on broad dimensions have been developed. In this report, we document the procedures and activities in developing MOS-specific performance appraisal forms that assess job effectiveness on broad behavioral dimensions. (Documentation of development activities of task-oriented performance measures may be found in Campbell, Campbell, Rumsey,& Edwards, 1986.)

This report contains three chapters. In Chapter 1, we describe the procedures used to develop behaviorally anchored performance rating scales, the sample of participants involved in defining the performance dimensions, and the resulting performance rating scales. Chapter 2 contains a description of the procedures used in field testing the newly developed scales, along

with results from the field test. Finally, in Chapter 3, we discuss decisions concerning rating scale modifications and present the final set of behaviorally anchored rating scales (BARS) to be used in the Concurrent Validation administration.

Background

The procedure used to identify MOS-specific job duties was derived in large part from procedures outlined by Smith and Kendall (1963) and by Campbell, Dunnette, Arvey, and Hellervik (1973). According to Smith and Kendall, performance appraisal rating scales should emphasize activity or performance that can be observed on the job. Their recommended procedure involves identifying behaviors that lead to effective or ineffective job performance outcomes and avoids focusing on unobservable or nonbehavioral attributes. Another feature of this methodology involves developing rating scales that incorporate the language of the users and that reflect standards which users help to define. Thus, activities to develop rating scales include the users in all phases of scale construction. Details of the development process are described below.

Smith and Kendall were the first to recommend using the critical incident technique described by Flanagan (1954) to identify the major dimensions or categories of job performance. This is accomplished by asking those most familiar with the job--supervisors and incumbents--to describe or write examples of effective, average, and ineffective behavior observed on the job.

These authors recommend conducting critical incident workshops that, as a first step, name and define the major components of performance for the job in question. Workshop participants are then asked to write examples of effective and ineffective performance for each of the major components they have identified.

Campbell et al. (1973) suggest a slight modification to the Smith and Kendall procedure. They recommend that performance categories be generated after participants have had an opportunity to write several incidents. In this way, participants will not be constrained by working with a priori performance categories and are more likely to write performance examples that represent all job requirements. Thus, it is less likely that important job duties will be overlooked.

The next step involves editing the written performance examples or critical incidents. Here, Smith and Kendall emphasize the need for retaining the "flavor" of the incidents to ensure that terminology used on the job also appears in the rating scales.

These edited incidents are then used to identify the major dimensions of the job. Two or more researchers independently content analyze the incidents and sort them into performance dimensions, and then compare their results to form a performance dimension system. Performance categories generated in workshop discussions may be used to help label and define the resulting performance dimensions.

Next, supervisors and incumbents are called in to participate in a retranslation exercise. They are asked to read the performance incidents and make two ratings for each. First, they must assign each incident to a performance dimension based on the behavior described in the incident. Second, raters are asked to indicate the effectiveness level of the behavior.

Results from this exercise are used to evaluate the performance dimension system to ensure that dimensions are clear and that raters can effectively allocate behavioral examples into each with a high level of agreement. Further, retranslation ratings are used to develop behavioral standards that represent performance at various effectiveness levels. The final product is a set of behaviorally defined and anchored performance dimensions that focus on the duties and standards of a specific job or MOS.

Guidelines for developing behaviorally anchored rating scales, established by Smith and Kendall (1963) and by Campbell et al. (1973), were used throughout the conduct of this part of Task 5. In the next section we describe in detail the development of behaviorally anchored rating scales for first-term enlistees.

Method

Target Military Occupational Specialties (MOS)

As noted, the purpose of this part of Task 5 was to develop behaviorally anchored performance rating scales that highlight specific job requirements for nine MOS. The pool of MOS that had been selected for inclusion in Project A comprised 19 specialties identified as representative of the more than 200 enlisted occupations in the Army.

Very early in the project it was deemed infeasible to develop specific job performance measurement instruments for all of the selected MOS. Therefore, a subset comprised of nine occupational specialties was selected for developing MOS-specific performance measures. These MOS were chosen on the basis of the total number of persons in each and the type of work performed. The objective was to identify MOS that have fairly large numbers and that represent different primary missions (i.e., combat arms, combat support, noncombat). The nine MOS selected are:

11B Infantryman
13B Cannon Crewman
19E Armor Crewman
31C Radio Teletype Operator (Originally coded 05C)
63B Light-Wheel Vehicle Mechanic
64C Motor Transport Operator
71L Administrative Specialist
91A Medical Specialist (Originally coded 91B)
95B Military Police

First, the nine MOS were divided into two groups or batches, Batch A and Batch B. The MOS in the first group (Batch A) are 13B, 64C, 71L, and 95B; those included in the second group (Batch B) are 11B, 19E, 31C, 63B, and

91A. Dividing the nine MOS into two groups made it possible to design and use data collection procedures for the first group, develop performance rating scales, and try them out in the field. Before beginning work on the second batch, we evaluated our procedures and modified them to improve and streamline the scale development process. For the most part, the procedures employed for the Batch A MOS are very similar to those used to develop scales for Batch B MOS. Where procedures differed for the two batches, we describe the differences and the rationale for the modifications.

Each of the nine MOS was assigned to a PDRI research staff member, who was responsible for (1) conducting workshops to collect performance incidents for the assigned MOS, (2) editing incidents, (3) preparing retranslation exercises, (4) developing performance rating scales, and (5) revising the scales for the Concurrent Validation efforts. Thus, a single researcher became an "expert" concerning the job duties and requirements involved in the assigned MOS.

Please note that we have prepared nine appendices that correspond to the nine MOS included in the project. These are located in a separate report, ARI Research Note ___, 1985 (four volumes). They appear in the following order: Appendix A - 13B Cannon Crewman; Appendix B - 64C Motor Transport Operator; Appendix C - 71L Administrative Specialist; Appendix D - 95B Military Police; Appendix E - 11B Infantryman; Appendix F - 19 E Armor Crewman; Appendix G - 31C Radio Teletype Operator; Appendix H - 63B Light-Wheel Vehicle Mechanic; and Appendix I - 91A Medical Specialist.

Sample

We modified the procedures somewhat from those described by Smith and Kendall (1963) and Campbell et al. (1973). For example, incumbents or first-term enlistees from target MOS were not, as a rule, included in the workshops. We reasoned here that first-termers, especially those who had been in the Army for only a year or two, would not have had the opportunity to obtain the "big picture" of MOS-specific job requirements. Therefore, to ensure that workshop participants were familiar with first-term enlistee job requirements, most individuals selected to participate in the workshops were non-commissioned officers (NCOs) directly responsible for supervising first-term enlistees and hence were equivalent to first-line supervisors. Further, most of the NCOs included in the sample had spent two to four years as first-termers in these MOS, and therefore were familiar with the job requirements from an "incumbent" as well as a "supervisor" perspective.

To ensure thorough coverage and representation of the critical behaviors in each MOS, workshops for each MOS were conducted at six CONUS (Continental United States) Army posts. Posts included in Batch A workshops were Fort Ord, California; Fort Polk, Louisiana; Fort Bragg, North Carolina; Fort Campbell, Kentucky; Fort Hood, Texas; and Fort Carson, Colorado. Those scheduled for Batch B workshops were Fort Lewis, Washington; Fort Stewart, Georgia; Fort Riley, Kansas; Fort Bragg, North Carolina; Fort Sill, Oklahoma; and Fort Bliss, Texas. The workshop schedule for collecting performance incidents at each of these sites is provided in Table 1.

Table 1

Workshop Locations and Dates

Location	Dates
Batch A	
Fort Ord	25 - 26 August 1983
Fort Polk	29 - 30 August 1983
Fort Bragg	12 - 13 September 1983
Fort Campbell	15 - 16 September 1983
Fort Hood	13 - 14 October 1983
Fort Carson	31 October - 1 November 1983
Batch B	
Fort Lewis	9 - 11 January 1984
Fort Stewart	11 - 13 January 1984
Fort Riley	16 - 18 January 1984
Fort Bragg	27 - 29 February 1984
Fort Bliss	12 - 14 March 1984
Fort Sill	14 - 16 March 1984

At each Army post, our point-of-contact (POC) was asked to obtain from 10 to 16 NCOs from each target MOS. Thus, the goal was to obtain input from about 60 to 96 supervisors for each MOS. The total numbers of NCOs participating in the performance incident workshops by MOS were as follows: 13B--N=88; 64C--N=81; 71L--N=63; 95B--N=86; 11B--N=83; 19E--N=65; 31C--N=60; 63B--N=75; and 91A--N=71.

A breakdown of each MOS workshop sample by rank and by gender is provided in Tables 2 and 3 for Batch A and Batch B MOS. For one MOS the total number of participants reported by rank does not equal the total reported above, because a few participants did not report their rank. It is also important to note that for three MOS no females participated, because these three MOS--13B, 19E, and 11B--involve combat duty, which precludes females from enlisting in them.

As the information in the tables indicates, the bulk of the workshop samples consisted of NCOs at the E-5 and E-6 levels. In some cases, however, participants were enlistees of lower rank, such as E-1 and E-2; these individuals were first-term enlistees with less that one year of job experience. Also, some workshop sessions contained NCOs at the E-8 and E-9 level. These individuals have less direct responsibilities for supervising first-term enlistees and can be considered equivalent to second-line supervisors.

Performance Incident Data Collection Activities

Workshop Description. We began each workshop session by providing participants with booklets containing information about Project A and about the day's activities. We have included the booklets used for each MOS in Section 1 of Appendices A through I.

The schedule of activities followed for each critical incident workshop for all MOS is shown in Table 4. Workshop leaders first provided a description of Project A, then briefed participants on the purpose of the workshop. This led to discussion of the different types of performance rating scales available, and the advantages of using behaviorally anchored rating scales to assess job performance. Leaders then described how the results from the day's activities would be used to develop this type of rating scale for that particular MOS.

Next, workshop leaders provided instruction for writing performance incidents. This included a description of the information required in each incident, such as the setting, the behaviors observed, and the outcome (or what happened as a result of the behavior). Participants were asked to review several examples in their booklets to get an idea of how to write performance incidents. The examples of "bad" incidents contained irrelevant information or lacked important information, whereas the "good" examples were corrected versions that contained all necessary information.

Workshop leaders then distributed performance incident forms and asked participants to generate performance incidents, using the examples as guides. Figure 1 shows a sample form that participants used to generate incidents.

Job Described _____

1. What were the circumstances leading up to the incident?

2. What did the individual do that made you feel he or she was a good, average, or poor performer?

3. In what job performance category would you say this incident falls?

4. Circle the number below that best reflects the correct effectiveness level for this example:

1	2	3	4	5	6	7	8	9
extremely ineffective		ineffective		about average		effective		extremely effective

Figure 1. Sample Performance Incident Form

Table 2

Performance Incident Workshops:

Rank and Gender of Batch A Participant Sample by MOS

13B - Cannon Crewman

Rank	N	%
E1	0	0.0
E2	0	0.0
E3	0	0.0
E4	2	2.3
E5	49	55.7
E6	29	33.0
E7	7	8.0
E8	1	1.1
E9	0	0.0
Total	88	
Gender		
M	88	100
F	0	0

64C - Motor Transport Operator

Rank	N	%
E1	0	0.0
E2	0	0.0
E3	3	3.9
E4	4	5.2
E5	34	44.7
E6	27	35.5
E7	8	10.5
E8	0	0.0
E9	0	0.0
Total	76	
Gender		
M	74	97.4
F	2	2.6

71L - Administrative Specialist[a]

Rank	N	%
E1	0	0.0
E2	1	1.6
E3	3	4.9
E4	0	0.0
E5	27	44.3
E6	10	16.4
E7	12	19.7
E8	7	11.5
E9	1	1.6
Total	61	
Gender		
M	44	69.8
F	19	30.2

95B - Military Police

Rank	N	%
E1	0	0.0
E2	0	0.0
E3	0	0.0
E4	0	0.0
E5	39	45.3
E6	24	27.9
E7	16	18.6
E8	6	6.9
E9	1	1.2
Total	86	
Gender		
M	84	97.7
F	2	2.3

[a]The total sample size by rank does not equal the total sample by gender because two individuals failed to report their rank.

Table 3

Performance Incident Workshops:

Rank and Gender of Batch B Participant Sample by MOS

11B - Infantryman

Rank	N	%
E1	0	0.0
E2	0	0.0
E3	6	7.3
E4	5	6.1
E5	32	39.0
E6	20	24.4
E7	13	15.9
E8	6	7.3
E9	0	0.0
Total	82	
Gender		
M	82	100
F	0	0

19E - Armor Crewman

Rank	N	%
E1	1	1.5
E2	0	0.0
E3	9	13.8
E4	12	18.5
E5	28	43.1
E6	13	20.0
E7	2	3.0
E8	0	0.0
E9	0	0.0
Total	65	
Gender		
M	65	100
F	0	0

31C - Radio Teletype Operator

Rank	N	%
E1	0	0.0
E2	2	3.3
E3	2	3.3
E4	4	6.7
E5	38	63.3
E6	14	23.3
E7	0	0.0
E8	0	0.0
E9	0	0.0
Total	60	
Gender		
M	52	86.7
F	8	13.3

Continued

Table 3 (Continued)

Performance Incident Workshops:

Rank and Gender of Batch B Participant Sample by MOS

63B - Light-Wheel Vehicle Mechanic			91A - Medical Specialist		
Rank	N	%	Rank	N	%
E1	1	1.3	E1	1	1.4
E2	3	4.0	E2	2	2.8
E3	4	5.3	E3	1	1.4
E4	5	6.7	E4	13	18.3
E5	35	46.7	E5	26	36.6
E6	20	26.7	E6	17	23.9
E7	6	8.0	E7	8	11.3
E8	1	1.3	E8	3	4.2
E9	0	0.0	E9	0	0.0
Total	75		Total	71	
Gender			Gender		
M	72	96.0	M	54	76.1
F	3	4.0	F	17	23.9

Table 4

Agenda for Performance Incident Workshop

Time	Topic
0800 - 0815	Description of the project
0815 - 0845	Briefing on the day's activities
0845 - 1130	Generating performance examples
1130 - 1230	Lunch
1230 - 1430	Generating more performance examples
1430 - 1530	Discussion of performance categories emerging in the workshop
1530 - 1615	Generating more performance examples
1615 - 1630	Review of the day's activities and discussion of the next steps

While writing performance incidents, participants were encouraged to avoid activities or behaviors that reflect general soldier effectiveness (e.g., following rules and regulations, military appearance); such requirements have been identified and described in a separate part of the project. (See Borman, Motowidlo, Rose & Hanser, 1984; and Borman & Rose, 1986 for a complete description of the Army-wide rating scales designed to assess general soldier effectiveness.)

As indicated earlier, the objective of these workshops was to generate examples of effective, average, and ineffective performance in each of the target MOS. To ensure thorough coverage of each MOS, workshop leaders established goals for participants. Participants were informed early in the day that each was expected to generate about 14 to 16 incidents; for the entire group, we requested about 200 performance incidents. (This goal applied to groups with 12 to 16 participants; it was modified accordingly for smaller groups.) To many participants that goal seemed unreasonably high, but as each workshop session progressed, it became clear that all participants could (and usually did) meet the established goals.

As participants finished writing an incident, workshop leaders reviewed it to ensure that it clearly described the situation, the behavior or activity, and the outcome of the incident. They also identified terminology and Army acronyms that were unclear or obscure and asked participants to clarify them.

Participants continued to generate performance incidents until it was time to break for lunch. Following lunch, workshop leaders asked participants to resume writing incidents for about two more hours. At that time, performance incident writing was halted and workshop leaders began generating discussion among participants to identify the major components or activities comprising the job or MOS.

During this discussion, participants were asked to identify the major job performance categories. Workshop leaders recorded suggested categories on a blackboard or flipchart. When participants indicated that all possible performance categories had been identified, the leader asked them to review the list and consider whether or not all job duties did indeed appear. The leader also asked them to consider whether each category represented first-term enlistee job requirements or requirements of more experienced soldiers.

Following this discussion, participants were asked to review the performance incidents they had written and to assign them to one of the job categories or dimensions that appeared on the blackboard or flipchart. The workshop leader then tallied the total number of incidents in each category. Those categories with very few incidents were the focus of the remainder of the workshop; participants were asked to spend the remaining time generating performance incidents for those categories represented by only a few performance incidents.

At the end of the session, workshop leaders discussed the next steps in the project. We informed participants that in a few months they would be asked to participate in another part of the study, which would involve retrans-

lating the performance incidents collected from all NCOs in the same MOS. The plan for this portion of the rating scale development strategy involved mailing the retranslation exercise to all participants. (This strategy was used only for Batch A MOS; for Batch B a slightly different approach was used.) Details about the retranslation exercise are provided later in this chapter.

Results from the performance incident workshops are reported in Table 5 for Batch A MOS and in Table 6 for Batch B MOS. In these tables, we report the number of workshop participants and number of performance incidents generated by MOS and by location, as well as the mean number of incidents generated by MOS and location. The tables also show the total number of participants and total number of incidents by MOS and by location.

For Batch A, the total number of participants for each MOS ranged from 63 for Administrative Specialist (71L) to 88 for the Cannon Crewman (13B) group. The number of incidents generated within each MOS ranges from 989 for the Administrative Specialist (71L) to 1183 for Military Police (95B). Finally, the average number of performance incidents provided by participants within MOS ranged from 13.2 for Cannon Crewman (13B) to 15.7 for Administrative Specialist (71L).

For Batch B, the total number of participants within MOS ranged from 60 for Radio Teletype Operator (31C) to 83 for Infantryman (11B). The total number of incidents generated for each MOS ranged from 761 for Medical Specialist (91A) to 993 for Infantryman (11B). (The total number of incidents generated within an MOS was less for Batch B MOS than for Batch A MOS, due to modifications in the procedures used for the Batch B retranslation exercise. These modifications are described in the Retranslation section of this chapter.) The average number of incidents generated by each participant within an MOS ranged from 10.7 for Medical Specialist (91A) to 13.0 for Radio Teletype Operator (31C).

These data indicate that we were successful in obtaining the number of participants requested, and that participants in each MOS provided an ample number of performance incidents for developing behaviorally anchored rating scales reflecting MOS-specific job requirements.

Activities Between Workshop Sessions. Performance incident workshops for each batch were conducted over a period of three months. This schedule permitted the research staff to edit and review performance incidents between data collection activities. Thus, for Batch A MOS, staff members edited incidents collected at Fort Ord and Fort Polk before collecting more incidents at Fort Bragg and Fort Campbell. Also during this time, staff members reviewed the incidents and the performance categories generated in the group discussion to construct a preliminary performance dimension system.

These performance dimensions were then presented and discussed at Fort Bragg and Fort Campbell. Following the data collection activities at these posts, the process was again repeated. That is, performance incidents were edited, content analyzed, and sorted into categories. These categories were then integrated with those generated during the discussion with workshop participants. And, once again, the new performance dimension catego-

Table 5

Performance Incident Workshops: Number of Participants and Number of Incidents Generated by MOS and by Location - Batch A

Location	MOS 13B	64C	71L	95B	Total By Location
Fort Ord					
N - Participants	14	10	5	14	43
N - Incidents	195	80	59	213	547
Mean Per Participant	13.9	8.0	11.8	15.2	12.7
Fort Polk					
N - Participants	12	15	15	15	57
N - Incidents	150	240	210	235	835
Mean Per Participant	12.5	16.0	14.0	15.7	14.7
Fort Bragg					
N - Participants	13	14	11	17	55
N - Incidents	235	221	218	225	899
Mean Per Participant	18.1	15.8	19.8	13.2	16.4
Fort Campbell					
N - Participants	17	14	9	15	55
N - Incidents	195	191	154	238	778
Mean Per Participant	11.5	13.6	17.1	15.9	14.2
Fort Hood					
N - Participants	13	13	10	11	47
N - Incidents	180	183	133	92	588
Mean Per Participant	13.9	14.1	13.3	8.4	10.7
Fort Carson					
N - Participants	19	15	13	14	61
N - Incidents	204	232	215	180	831
Mean Per Participant	10.7	15.5	16.5	12.9	13.6
Totals By MOS					
N - Participants	88	81	63	86	318
N - Incidents	1159	1147	989	1183	4478
Mean Per Participant	13.2	14.2	15.7	13.8	14.1

Table 6

Performance Incident Workshops: Number of Participants and Number of Incidents Generated by MOS and by Location - Batch B

Location	MOS 11B	19E	31C	63B	91A	Total by Location
Fort Lewis						
N - Participants	16	11	8	10	11	56
N - Incidents	211	180	124	172	130	817
Mean Per Participant	13.8	16.4	15.5	17.2	11.8	14.6
Fort Stewart						
N - Participants	14	15	15	16	16	76
N - Incidents	216	275	256	208	249	1204
Mean Per Participant	15.4	18.3	17.1	13.0	15.6	15.8
Fort Riley						
N - Participants	18	7	10	11	8	54
N - Incidents	216	123	127	133	90	689
Mean Per Participant	12.0	17.6	12.7	12.1	11.3	13.8
Fort Bragg						
N - Participants	13	14	16	15	13	71
N - Incidents	231	190	220	250	217	1108
Mean Per Participant	17.8	13.6	13.8	16.7	16.7	15.6
Fort Sill[a]						
N - Participants	8	4	3	9	10	34
N - Incidents	26	0	13	32	20	91
Mean Per Participant	3.3		4.3	3.6	2.0	2.7
Fort Bliss[a]						
N - Participants	14	14	8	14	13	63
N - Incidents	93	70	39	71	55	328
Mean Per Participant	6.6	5.0	4.9	5.1	4.2	5.2
Total by MOS						
N - Participants	83	65	60	75	71	354
N - Incidents	993	838	779	866	761	4237
Mean Per Participant	12.0	12.9	13.0	11.6	10.7	12.0

[a]Participants at these posts spent most of the time completing retranslation booklets rather than generating critical incidents.

ries were presented and discussed with participants in workshops held at Fort Hood and Fort Carson.

A similar iterative procedure was used to generate Batch B performance dimensions. Performance incidents collected at Fort Lewis, Fort Stewart, and Fort Riley were edited, content analyzed, and then sorted into performance dimensions. Results from the sort were presented and discussed at the next site, Fort Bragg. The procedures followed for the final two forts for Batch B, Fort Sill and Fort Bliss, differed slightly from those used for Batch A MOS; these procedural differences are discussed in the next section.

Retranslation Activities

Rationale. A primary purpose of the retranslation exercise is to verify that the performance dimension system represents thorough and comprehensive coverage of the critical job requirements. Persons familiar with the target job are asked to review the performance incidents generated for that job.

After reviewing each incident, participants must first assign it to one of the performance dimensions. The objective here is to identify performance incidents with high levels of agreement (e.g., 50% or greater) in performance dimension assignment.

A second objective is to construct performance anchors for each dimension. This information is obtained from a second rating participants provide for each incident, which involves evaluating the effectiveness of the behavior described. These ratings are used to help define each performance dimension and to construct behavioral anchors that describe typical performance at different effectiveness levels within that dimension. Such anchors are designed to ensure that raters use the same standards of performance to evaluate ratees. That is, they provide raters with systematic information about behaviors that comprise ineffective performance, average performance, and effective performance within a particular dimension.

Performance dimension anchors are derived directly from performance incidents. To construct anchors, performance incidents that all or most raters agree describe activity in a single performance dimension are identified along with incidents that most raters agree depict performance at a particular effectiveness level. Those incidents are then used to develop the anchor for performance at that effectiveness level. In summary, we are looking for high agreement among raters on performance dimension assignment of incidents (or high percentage agreement) and high agreement among raters for the effectiveness level demonstrated in each incident (or low standard deviations).

Retranslation procedures employed for Batch A MOS differed from those for Batch B MOS. Below we describe the activities in retranslating the performance incidents for Batch A MOS. We then discuss some of the problems in using these procedures and the modifications made for Batch B MOS retranslation activities.

Retranslation Materials and Procedures - Batch A. The Smith and Kendall (1963) procedure calls for including individuals familiar with the target job to participate in the retranslation process. For the Batch A MOS, we planned to include workshop participants in this phase of the project. (Recall that these persons were supervisors of the target incumbents and, hence, as a rule, did not include the incumbent group.) During the performance incident workshops participants were informed that we would contact them via the mail to complete another phase of project.

In the last performance incident workshop, conducted at Fort Carson, participants for each MOS were given a "practice" retranslation package which included instructions for completing the exercise, a list and description of performance dimensions, and a subset of the edited performance incidents. The number of incidents retranslated varied by MOS; 13B examined 240 incidents, 64C 14 incidents, 71L up to 200 incidents, and 95B 100 incidents.

This "practice" retranslation exercise was conducted to ensure that the instructions and completed example incidents clearly explained the task. Workshop leaders simply passed out the materials to participants and instructed them to complete the task; no further instructions were provided. As participants finished, leaders noted any questions or problems that they had experienced. This information was used to modify the retranslation instructions and the example items. The final sets of retranslation materials, including instructions, examples, and performance dimensions and definitions, are provided in Section 2 of the MOS appendices.

In designing the retranslation exercise booklets, we first screened all performance incidents and removed duplicates, incidents that were unclear or incomplete, and any that depicted Army-wide rather than MOS-specific job requirements.

After taking a count of the remaining incidents, we concluded that it was impractical to ask participants to rate all performance incidents generated for their MOS. As shown in Tables 5 and 6, the number of incidents generated for each MOS ranged from 761 to 1183 (the actual number of performance incidents was somewhat lower than that due to the screening procedures employed). Instead, we constructed a less onerous task that asked participants to retranslate only a subset of the total number; they were asked, on the average, to retranslate about 200 performance incidents. Thus, for each MOS we constructed four or five booklets containing unique performance incidents for the retranslation exercise.

Return rates across all Batch A MOS indicated that, on the average, only about 20 percent of the participants completed the retranslation task. This number proved insufficient for the analyses we planned. To increase the number of retranslation ratings, we conducted retranslation workshops at Fort Meade, Maryland. These workshops included NCOs from the four MOS who were familiar with first-term enlistee job requirements. Project staff members from HUMRRO who were familiar with the job requirements of one or more MOS also completed retranslation booklets.

Procedures for Batch B. Because of the low return rate from mailing out retranslation materials for Batch A, we modified the procedures for obtain-

ing retranslation ratings for the Batch B MOS. Non-commissioned officers from six locations were asked to participate in the Batch B performance incident workshops. The first four workshops were conducted in the same manner as those for Batch A MOS; participants spent a majority of their time generating incidents, with an hour or two spent discussing the critical performance categories comprising the job. At the final two workshops, conducted at Fort Sill and Fort Bliss, participants spent the first two hours generating performance incidents describing MOS-specific job behaviors, then spent the remainder of their day completing retranslation booklets.

Retranslation materials administered in these sessions were very similar to those administered to Batch A participants. That is, for each MOS we constructed retranslation booklets that contained about 200 to 270 performance incidents. Thus, retranslation materials for each Batch B MOS included from two to three booklets that contained unique performance incidents. (Retranslation materials administered to Batch B MOS appear in Section 2 of the separate appendices.)

During the final two workshop sessions, we asked participants to complete as many retranslation booklets as possible. In general, participants completed about one-and-one-half to two booklets. Also during this session, participants were asked to retranslate the performance incidents generated earlier during that session. Hence, we obtained retranslation ratings for all performance incidents generated at the first four workshops and for the new incidents generated at that particular workshop.

Results from Retranslation Ratings

Table 7 summarizes the number of ratings obtained from the retranslation exercise for Batch A and Batch B. This table indicates again that we obtained a greater number of incidents for Batch A MOS than for Batch B MOS. The average number of ratings per retranslation booklet varied for the nine MOS, ranging from 7.6 for Military Police (95B) to 19.0 for Infantryman (11B). In general, we obtained about nine or ten ratings for each performance incident contained in the retranslation exercise.

As noted above, individuals completing the retranslation exercise were asked to read each performance incident and provide two ratings: (1) assign the incident to a performance dimension based on the behavior depicted in the incident, and (2) rate the effectiveness of the behavior using a scale of 1 for ineffective performance to 9 for effective performance (a value of 5 on this scale represents average performance).

Analysis of the retranslation data was conducted separately for each MOS. This included computing for each incident: (1) the number of raters; (2) percent agreement among raters in assigning incidents to performance dimensions; (3) mean effectiveness rating; and (4) standard deviation of the effectiveness ratings. Percent agreement values, mean effectiveness ratings, and standard deviations are provided for all performance incidents included in the retranslation exercise in Section 3 of the MOS appendices.

Table 7

Retranslation Exercise: Number of Forms Developed for Each MOS and Average Number of Raters Completing Each Form

MOS	Number of Forms	Average Number of Incidents/Form (Total Number of Incidents)	Average Number of Raters/Form
Batch A			
13B	4	171 (684)	17.0
64C	5	191 (955)	12.6
71L	4	190 (760)	14.0
95B	5	229 (1145)	7.6
Batch B			
11B	2	274 (548)	19.0
19E	3	201 (603)	9.7
31C	3	235 (705)	9.0
63B	3	230 (690)	16.0
91A	3	210 (630)	14.7

Development of Behaviorally Anchored Rating Scales

The next step in the process involved identifying those performance incidents in which raters agreed fairly well on performance dimension assignment and effectiveness level. For each MOS, we identified performance incidents that met the following criteria: (1) at least 50% of the raters agreed that the incident depicted performance in a single performance dimension; and (2) the standard deviation of the mean effectiveness rating did not exceed 2.0.

We then sorted these incidents into their assigned performance dimensions. Results from this sorting are presented for each MOS in Tables 8 through 16 and are discussed in detail in the next section of this chapter. The performance dimensions listed in these tables were the ones used by raters in the retranslation exercise; they do not necessarily reflect the performance dimensions administered in the field test sessions described in Chapter 2.

After all incidents had been sorted into performance dimensions, we examined the incidents and the percentage agreement values in each dimension. Recall that previously we had identified all performance incidents for which at least 50% of the raters agreed in dimension assignment. We carefully reviewed those incidents with percentage agreement at the 50% level to identify performance dimensions that raters found confusing or difficult to distinguish one from another. For example, most raters for the Armor Crewman (19E) MOS agreed that incidents describing tank hull or tank turret system maintenance should be assigned to either "Maintaining tank/hull suspension system and associated equipment" (Dimension A) or "Maintaining tank turret/fire control system" (Dimension B) (see Table 13). It appeared that tank maintenance activities could not be clearly distinguished by tank component, so these two performance dimensions were combined into one.

After evaluating our performance dimension systems and modifying them using results from the retranslation exercise, we began developing behavioral anchors for each dimension. This involved sorting performance incidents into three effectiveness-level categories--effective performance with mean values of 6.5 or higher, average performance with mean values of 3.5 to 6.4, and ineffective performance with mean values of 1.0 to 3.4. We reviewed the content of the incidents in each of these three areas and then summarized the information in each to form three behavioral anchors depicting effective, average, and ineffective performance.

It is important to note that for each MOS we developed Behavioral Summary Scales. Traditional behaviorally anchored rating scales contain specific examples of job behaviors for each effectiveness level in a performance dimension. Behavioral Summary Scales, on the other hand, contain anchors that represent the behavioral content of <u>all</u> performance incidents reliably retranslated for that particular level of effectiveness. This makes it more likely that a rater using the scales will be able to match observed performance with performance on the rating scale (Borman, 1979). A sample of one behavioral summary scale constructed for one MOS, Military Police (95B), is presented in Figure 2.

A. TRAFFIC CONTROL AND ENFORCEMENT

Controlling traffic and enforcing traffic laws and parking rules.

1 2	3 4 5	6 7
• Often uses hand/arm signals that are difficult to understand, at times resulting in unnecessary accidents; often fails to wear reflectorized gear; overlooks hazardous traffic conditions; sleeps on duty; pays excessive attention to things unrelated to the job.	• Usually does a reasonable job when directing traffic by using adequate hand/arm signals and/or wearing reflectorized gear.	• Consistently uses appropriate hand/arm signals; always wears reflectorized gear; generally monitors traffic from plain-view vantage points; consistently refrains from behaviors such as reading and prolonged conversation on non-job related topics.
• May display excess leniency or harshness when citing offenders, allowing their military rank, race, and/or sex to influence his/her actions; makes many errors when filling out citations.	• Makes few errors when filling out citations; usually does not allow an offender's race, sex, and/or military rank to interfere with good judgment.	• Always uses emergency equipment (e.g., flares, barricades) to highlight unsafe conditions and ensures that hazards are removed or otherwise taken care of.

Figure 2. Sample Behavioral Summary Rating Scale for Military Police (95B)

It is evident from Tables 8 through 16 that some performance dimensions contained a small number of reliably sorted incidents. When this occurred, we reconsidered including that performance dimension in the rating scales. For some MOS, these dimensions were omitted or, where appropriate, combined with another performance dimension. To combine these dimensions with other dimensions, we examined the percentage agreement values to determine whether or not raters confused the dimension in question with another performance dimension. In some cases, we retained the performance dimension because it represented requirements that, although performed infrequently, are critical for success on the job. Behavioral anchors for such dimensions were developed by extrapolating information from available performance incidents.

After developing the performance rating scales for each MOS, we submitted the scales for review, generally by a PDRI research staff member familiar with the development process. Results from this review were used to clarify performance definitions and behavioral anchors. The final set of performance rating scales administered in field test sessions are included in the MOS appendices, Section 4.

Results and Revisions

Below we describe results from the retranslation data for each MOS and the modifications made to the scales.

Cannon Crewman (13B). For the retranslation exercise, 10 performance dimensions were identified from the performance incidents collected. Results from the retranslation exercise indicate that the number of incidents reliably sorted into these dimensions ranged from 14 to 195 (see Table 8). Most incidents appeared for "Driving and maintaining vehicles, Howitzers, and equipment" (Dimension B) and "Transporting/ sorting/storing and preparing ammunition for fire" (Dimension C). Although only a small number of incidents were reliably sorted into "Receiving and relaying communications" (Dimension H) and "Position improvement" (Dimension J), these dimensions were retained because they represent important activities in the Cannon Crewman MOS.

The final set of rating scales contains all of the ten original performance dimensions. They appear as follows: A. Loading out equipment; B. Driving and maintaining vehicles, Howitzers, and equipment; C. Transporting/sorting/ storing and preparing ammunition for fire; D. Preparing for occupation/ emplacing Howitzer; E. Setting up communications; F. Gunnery; G. Loading/ unloading Howitzer; H. Receiving and relaying communications; I. Recording/ record keeping; and J. Position improvement. (See Appendix A, Section 4 for complete scale definitions and anchors.)

Motor Transport Operator (64C). A sorting of the performance incidents revealed that 10 dimensions described the job requirements for this MOS. The number of incidents reliably sorted into each dimension ranged from 15 to 181 (see Table 9). Dimensions containing the largest number of reliably sorted incidents include "Checking and maintaining vehicles" (Dimension C) and "Driving vehicles" (Dimension A). Although one dimension, "Performing dispatcher duties" (Dimension J), contains a small number of incidents,

Table 8

Cannon Crewman (13B): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Loading out equipment	49
B. Driving and maintaining vehicles, Howitzers, and equipment	195
C. Transporting/sorting/storing and preparing ammunition for fire	108
D. Preparing for occupation and emplacing Howitzer	44
E. Setting up communications	24
F. Gunnery	99
G. Loading/unloading Howitzer	32
H. Receiving and relaying communications	19
I. Recording/record keeping	29
J. Position improvement	14
Total Number	613

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

Table 9

Motor Transport Operator (64C): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Driving vehicles	158
B. Vehicle coupling	46
C. Checking and maintaining vehicles	181
D. Using maps/following proper routes	27
E. Loading cargo and transporting personnel	75
F. Parking and securing vehicles	32
G. Performing administrative duties	42
H. Self-recovering vehicles	20
I. Safety-mindedness	80
J. Performing dispatcher duties	15
Total Number	676

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

this was retained because it represents an important requirement of the Motor Transport Operator position.

The final set of 10 rating scales includes: A. Driving vehicles; B. Vehicle coupling; C. Checking and maintaining vehicles; D. Using maps/following proper routes; E. Loading cargo and transporting personnel; F. Parking and securing vehicles; G. Performing administrative duties; H. Self-recovering vehicles; I. Safety-mindedness; and J. Performing dispatcher duties. (See Appendix B, Section 4 for complete scale definitions and anchors.)

Administrative Specialist (71L). For the retranslation exercise, we derived 13 performance dimensions from a sorting of the performance incidents. The number of incidents reliably sorted into each ranged from 2 to 183 (see Table 10). Dimensions containing the largest number of incidents include "Preparing, typing, and proofreading documents" (Dimension A) and "Keeping records" (Dimension F).

We modified the performance dimension system after reviewing the retranslation results. First, we decided to drop Dimensions I through M. "Preparing special reports, document drafts, or other materials" (Dimension I) was deleted because it described skills and activities more frequently performed by only the most experienced first-termers and by second-termers. Dimensions J through M were omitted because they involve job requirements for a subset of incumbents within the 71L position--71L F5 or Postal Clerk. These dimensions were identified very early in the workshop sessions and we encouraged participants to generate behavioral examples of these activities, when possible. It is clear from the retranslation data, however, that very few participants generated examples describing these duties and/or very few incidents were reliably sorted into these performance categories. Therefore, we decided to omit these dimensions.

The final set of Administrative Specialist rating scales includes: A. Preparing, typing, and proofreading documents; B. Distributing and dispatching incoming and outgoing documents; C. Maintaining office resources; D. Posting regulations; E. Establishing and/or maintaining files IAW TAFFS; F. Keeping records; G. Safeguarding and monitoring security of classified documents; and H. Providing customer service. (See Appendix C, Section 4 for complete scale definitions and anchors.)

Military Police (95B). A content analysis of the performance incidents revealed that seven dimensions effectively represented the requirements for this MOS. The number of incidents reliably sorted into these dimensions ranged from 50 to 236 (see Table 11). Dimensions containing the largest number of incidents are "Patrolling and crime/accident prevention activities" (Dimension D) and "Making arrests, gathering information on criminal activity, and reporting on crimes" (Dimension C).

We modified the performance dimensions only slightly; we shortened dimension titles. The final set of performance dimensions appears as follows: A. Traffic control and enforcement; B. Providing security; C. Investigating crimes and making arrests; D. Patrolling; E. Promoting the public image of the Military Police; F. Interpersonal communication skills; and G. Responding to medical emergencies. (See Appendix D, Section 4 for complete scale definitions and anchors.)

Table 10

Administrative Specialist (71L): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Preparing, typing, and proofreading documents	183
B. Distributing and dispatching incoming/outgoing documents	63
C. Maintaining office resources	73
D. Posting regulations	44
E. Establishing and/or maintaining files IAW TAFFS	50
F. Keeping records	94
G. Safeguarding and monitoring security of classified documents	43
H. Providing customer service	30
I. Preparing special reports, document drafts, or other materials	19
J. Sorting, routing and distributing incoming/outgoing mail	28
K. Maintaining Army Post Office equipment	2
L. Keeping Post Office records	20
M. Maintaining security of mail	9
Total Number	658

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

Table 11

Military Police (95B): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Traffic control and enforcement on post and in the field	63
B. Providing escort security and physical security	128
C. Making arrests, gathering information on criminal activity, and reporting on crimes	173
D. Patrolling and crime/accident prevention activities	236
E. Promoting confidence in the military police by maintaining personal and legal standards and through community service work	118
F. Using interpersonal communication (IPC) skills	87
G. Responding to medical emergencies and other emergencies of a non-criminal nature	50
Total Number	855

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

Infantryman (11B). For the retranslation exercise, 13 performance dimensions were identified through a content analysis of the performance incidents. Results from this exercise revealed that raters reliably sorted from 5 to 91 incidents into each performance dimension (see Table 12). The greatest numbers of incidents were reliably sorted into "Demonstrating proficiency in the use of all weapons, armaments, equipment, and supplies" (Dimension E) and in "Perform guard and security duties" (Dimension K).

An examination of the percent agreement values indicated that raters frequently confused "Using weapons safely" (Dimension D) and "Demonstrating proficiency in the use of all weapons, armaments, equipment, and supplies" (Dimension E). Therefore, we decided to combine these two to form a single dimension, "Use of weapons and other equipment."

We decided to retain one of the dimensions that contained only a few performance incidents, "Demonstrating courage and proficiency in engaging the enemy" (Dimension L), because it represented a critical Infantryman activity.

The only modification made to the remaining performance dimensions involved renaming them; virtually all dimensions received new titles. We labeled the final set of 12 dimensions as follows: A. Maintaining supplies, equipment, and weapons; B. Assisting and leading others; C. Navigation; D. Use of weapons and other equipment; E. Field sanitation, personal hygiene, and safety; F. Fighting position; G. Avoiding enemy detection; H. Operating a radio; I. Reconnaissance and patrol; J. Guard and security duties; K. Courage and proficiency in battle; and L. Prisoners of war. (See Appendix E, Section 4 for complete scale definitions and anchors.)

Armor Crewman (19E). A content analysis of the performance incidents revealed that 11 performance dimensions described the major components of the Armor Crewman job (see Table 13). Retranslation raters reliably sorted from 11 to 123 incidents into each dimension. The largest numbers of incidents appeared in "Maintaining tank, hull/suspension system and associated equipment" (Dimension A) and "Driving/recovering tanks" (Dimension C).

We modified the performance dimension system using results from the retranslation exercise. First, agreement values for Dimensions A and B indicated that raters frequently confused these two. Therefore, we decided to combine the two to form a single dimension, "Maintaining tank, tank systems, and associated equipment." For similar reasons "Establishing security in the field" (Dimension I) and "Preparing/securing tanks" (Dimension K) were combined to form a single dimension, "Preparing tanks for field problems." Finally, we decided to omit "Navigating" (Dimension J), because it contained only a few incidents and because this dimension appeared to represent job responsibilities required of more experienced or higher ranking soldiers.

The final set of rating scales contains 8 performance dimensions. These include: A. Maintaining tank, tank systems and associated equipment; B. Driving/recovering tanks; C. Stowing ammunition aboard tanks; D. Loading/unloading guns; E. Maintaining guns; F. Engaging targets with tank

Table 12

Infantryman (11B): Number of Behavioral Examples

Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Ensuring that all supplies and equipment are field-ready and available and well-maintained in the field	73
B. Providing leadership and/or taking charge in combat situations	33
C. Navigating and surviving in the field	53
D. Using weapons safely	38
E. Demonstrating proficiency in the use of all weapons, armaments, equipment, and supplies	91
F. Maintaining sanitary conditions, personal hygiene, and personal safety in the field	24
G. Preparing a fighting position	29
H. Avoiding enemy detection during movement and in established defensive positions	22
I. Operating a radio	27
J. Performing reconnaissance and patrol activities	37
K. Performing guard and security duties	75
L. Demonstrating courage and proficiency in engaging the enemy	5
M. Guarding and processing POWs and enemy casualties	15
Total Number	522

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

Table 13

Armor Crewman (19E): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Maintaining tank hull/suspension system and associated equipment	123
B. Maintaining tank turret system/fire control system	37
C. Driving/recovering tanks	80
D. Stowing and handling ammunition	39
E. Loading/unloading guns	30
F. Maintaining guns	43
G. Engaging targets with tank guns	45
H. Operating and maintaining communication equipment	36
I. Establishing security in the field	33
J. Navigating	11
K. Preparing/securing tank	27
Total Number	504

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

guns; G. Operating and maintaining communications equipment; and H. Preparing tanks for field problems. (See Appendix F, Section 4 for complete scale definitions and anchors.)

Radio Teletype Operator (31C). Initially, we identified seven performance dimensions to represent the job requirements for this MOS. Results from the retranslation exercise indicate that raters reliably sorted from 33 to 162 incidents into each dimension (see Table 14). The greatest numbers of incidents appeared in "Installing and preparing equipment for operation" (Dimension C) and "Operating communications devices and providing for an accurate and timely flow of information" (Dimension D).

We made one change in the performance dimension system. Results from the retranslation exercise indicated that raters frequently confused two of the dimensions, "Inspecting equipment and troubleshooting problems" (Dimension A) and "Pulling preventative maintenance and servicing equipment" (Dimension B). Hence, we combined these two into a single dimension, "Inspecting and servicing equipment." In addition, we renamed some of the performance dimensions.

The final set of rating scales contains the following six performance dimensions: A. Inspecting and servicing equipment; B. Installing and repairing equipment; C. Operating communications devices; D. Preparing reports; E. Maintaining security; and F. Providing safe transportation. (See Appendix G, Section 4 for complete scale definitions and anchors.)

Light-Wheel Vehicle Mechanic (63B). For the retranslation exercise, we identified 11 performance dimensions that represent the important requirements of the mechanic position. Retranslation raters reliably sorted from 15 to 101 incidents into each dimension, with the greatest numbers appearing in "Repair" (Dimension D), and "Safety-mindedness" (Dimension K) (see Table 15).

Performance rating scales developed for the field test included all 11 original dimensions. We reasoned that although "Vehicle and equipment operation" (Dimension G) and "Planning/organizing jobs" (Dimension I) contained a small number of incidents, these activities represented important components of the mechanic position. The only modification made to the scales involved reordering the final four dimensions. The final set of performance dimensions appears as follows: A. Inspecting and testing problems with equipment; B. Troubleshooting; C. Performing routine maintenance; D. Repair; E. Using tools and test equipment; F. Using technical documents; G. Vehicle and equipment operation; H. Safety mindedness; I. Administrative duties; J. Planning and organizing jobs; and K. Recovery. (See Appendix H, Section 4 for complete scale definitions and anchors.)

Medical Specialist (91A). The original system contained 11 performance dimensions. The number of incidents reliably sorted into each dimension ranged from 11 to 142 (see Table 16). The greatest numbers of incidents appeared in "Responding to emergency situations" (Dimension J), and "Providing routine and ongoing patient care" (Dimension I).

Modifications for the field test included deleting two performance dimensions. We omitted one dimension, "Attending to patient's concerns" (Dimen-

Table 14

Radio Teletype Operator (31C): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Inspecting equipment and troubleshooting problems	50
B. Pulling preventative maintenance and servicing equipment	79
C. Installing and preparing equipment for operation	162
D. Operating communications devices and providing for an accurate and timely flow of information	147
E. Preparing reports	33
F. Maintaining security of equipment and information	57
G. Locating and providing safe transport of equipment to sites	50
Total Number	578

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

Table 15

Light-Wheel Vehicle Mechanic (63B): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Inspecting, testing, and detecting problems with equipment	47
B. Troubleshooting	63
C. Performing routine maintenance	23
D. Repair	101
E. Using tools and test equipment	68
F. Using technical documentation	56
G. Vehicle and equipment operation	18
H. Recovery	36
I. Planning/organizing jobs	15
J. Administrative duties	41
K. Safety mindedness	89
Total Number	557

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

Table 16

Medical Specialist (91A): Number of Behavioral Examples Reliably Retranslated Into Each Dimension[a]

Dimension	Number of Examples
A. Maintaining and operating Army vehicles	51
B. Maintaining accountability of medical supplies and equipment	28
C. Keeping medical records	31
D. Attending to patients' concerns	15
E. Providing accurate diagnoses in a clinic, hospital, or field setting	11
F. Arranging for transportation and/or transporting injured personnel	44
G. Dispensing medications	42
H. Preparing and inspecting field site or clinic facilities in the field	34
I. Providing routine and ongoing patient care	95
J. Responding to emergency situations	142
K. Providing instruction to Army personnel	18
Total Number	511

[a]Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

sion D), because this particular activity appeared important for success in many of the performance dimensions. A second dimension, "Providing accurate diagnosis in a clinic, hospital, or field setting" (Dimension E), was omitted because it represented duties required of more experienced or higher ranking soldiers.

The final set of rating scales contains nine performance dimensions. These include: A. Maintaining and operating Army medical vehicles and equipment; B. Maintaining accountability of medical supplies and equipment; C. Keeping medical records; D. Arranging transportation and/or transporting injured personnel; E. Dispensing medications; F. Preparing and inspecting field site or clinic facilities; G. Providing routine and ongoing patient care; H. Responding to emergency situations; and I. Providing health care and health maintenance instruction to Army personnel. (See Appendix I, Section 4 for complete scale definitions and anchors.)

Preparation for Field Test

In sum, we relied on results from the retranslation exercise to evaluate and modify the performance dimension system for each MOS. Further, we generated behavioral anchors for each of the performance dimensions using results from our analysis of the retranslation ratings.

The final set of behaviorally anchored rating scales for the nine MOS, as described in the preceding section, contains from 6 to 12 performance dimensions. Each of the performance dimensions includes behavioral anchors describing ineffective, average, and effective performance. Raters are asked to use these anchors to evaluate ratees on a seven-point rating scale ranging from 1 (ineffective performance) to 7 (effective performance).

Before administering the rating scales in the field test, we constructed one additional rating scale for each MOS rating booklet. This scale asks raters to evaluate an incumbent's overall performance across all MOS-specific performance dimensions. This final rating scale is virtually the same for all MOS; it includes three anchors depicting ineffective, average, and effective performance.

Finally, we constructed rating scale booklets for each MOS that provided raters with performance dimension titles, definitions, and behavioral anchors. We designed rating booklets such that raters could evaluate up to five ratees in each. The booklets themselves do not include instructions for using the scales to make performance ratings. Our plan was to provide oral instructions during the field test rating sessions.

The MOS-specific rating scale booklets ask raters to evaluate incumbents on several performance dimensions specific to the target MOS job requirements and then to consider the incumbents' performance across all MOS-specific performance dimensions to arrive at an overall evaluation.

CHAPTER 2: MOS-SPECIFIC BEHAVIORALLY ANCHORED RATING SCALES: FIELD TEST ADMINISTRATION AND RESULTS

Introduction

Field test sessions were conducted separately for Batch A and Batch B MOS. We administered rating scales to Batch A MOS during the period of May through August 1984. These sessions were conducted at three CONUS sites and at two OCONUS (Outside Continental United States) sites. These included Fort Hood, Texas; Fort Polk, Louisiana; Fort Riley, Kansas; and two USAREUR sites (U.S. military posts located in West Germany).

Rating scales for Batch B MOS were field tested during the period of February through April 1985. Sessions were conducted at four CONUS locations and several OCONUS locations. These included Fort Lewis, Washington; Fort Polk, Louisiana; Fort Riley, Kansas; Fort Stewart, Georgia; and USAREUR locations in West Germany.

Administration procedures for the rating sessions were virtually the same for the two batches. Before describing those procedures, we describe the field test set-up to provide the context in which the rating scales were administered.

At each field test site, project staff administered several job performance and training performance measures to first-term enlistees. These measures were divided into four blocks: (1) hands-on tests of critical job tasks; (2) written job knowledge tests of critical tasks; (3) rating scales measuring performance in critical task areas both Army-wide and MOS-specific, and performance on broad behavioral dimensions both Army-wide and MOS-specific; and (4) written tests assessing knowledge acquired in Advanced Individual Training (AIT). The objective was to evaluate <u>all</u> training and performance measures that had been developed for Project A. Each type of measure was administered in a four-hour period. Thus, first-term enlistees participating in the field test sessions were scheduled to appear for two consecutive days.

The general plan for administering the four types of performance measures included scheduling 60 recruits from a particular MOS for the two-day period. This group was then divided into four smaller groups of fifteen. Over the two day period we rotated the four groups into the four job performance/training outcome assessment blocks. For example, Group A began by completing the hands-on test and then attended the rating session on Day One; on Day Two, Group A attended the written job knowledge test session in the morning and the written training knowledge test in the afternoon. Group B began with the written training knowledge test and the hands-on test on Day One; on Day Two, this group attended the rating session in the morning and completed the written job knowledge test in afternoon. Group C began with the written job knowledge test and then the written training knowledge test on Day One; Day Two activities included the hands-on test and then the rating session. Finally, Group D began with the written training knowledge test and then attended the rating session; on Day Two,

this group completed the job knowledge and hands-on tests. Figure 3 contains a sample schedule for one MOS at one test site location, USAREUR-Batch B.

The procedure described above was modified to accommodate soldiers from two MOS attending the field test session over the same two day period. In this case, we scheduled 30 soldiers from each MOS and again divided them into four groups of fifteen. The four groups completed the four performance measurement sessions on a rotational schedule. Figure 4 provides a sample schedule for a field test session that includes two different MOS for the same two-day period.

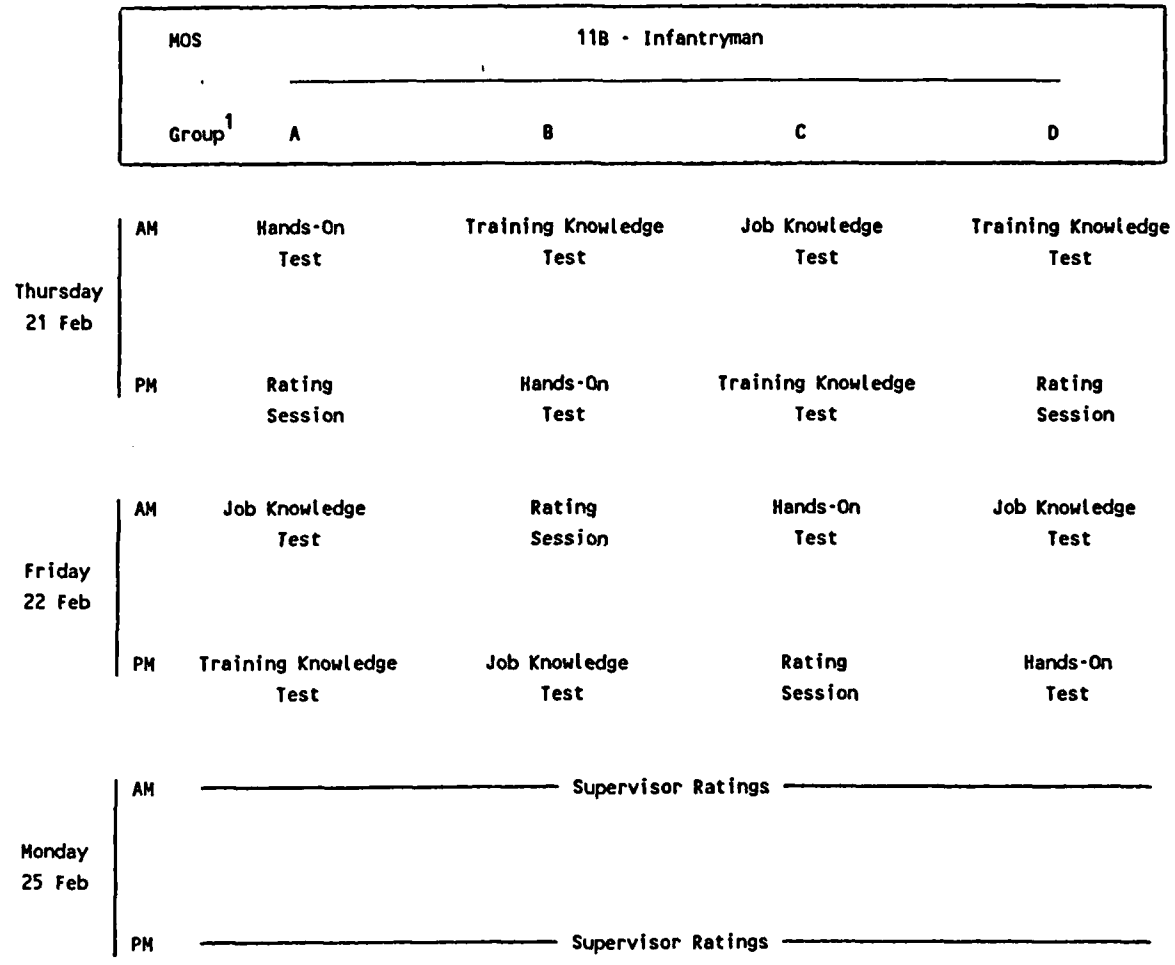
Our objective for all performance assessment sessions was to have administrators work closely with participants to ensure that everyone understood the instructions and to uncover any problems with the materials and the procedures. Specifically, for the rating sessions, we wanted to uncover any problems with the scales (e.g., whether raters understand the instructions for completing the rating scales, whether raters understand the performance dimensions and are able to use each to evaluate ratees' performance, what type of rater training is useful in this setting).

In the next section, we describe each sample participating in the field test sessions (by MOS), and then describe the procedures used to administer the rating scales. To present the context in which the MOS-specific behaviorally anchored rating scales (BARS) were administered, we describe the materials included in each rating session, and the rater training procedures. Our focus throughout this report is, however, on the MOS-Specific BARS, so in the results and discussion section, we deal exclusively with those scales. (Campbell et al., 1986, document development activities and field test results for hands-on measures and written job knowledge measures. Davis, Davis & Joyner, 1985, document development activities and field test results for job relevant training measures.)

Method

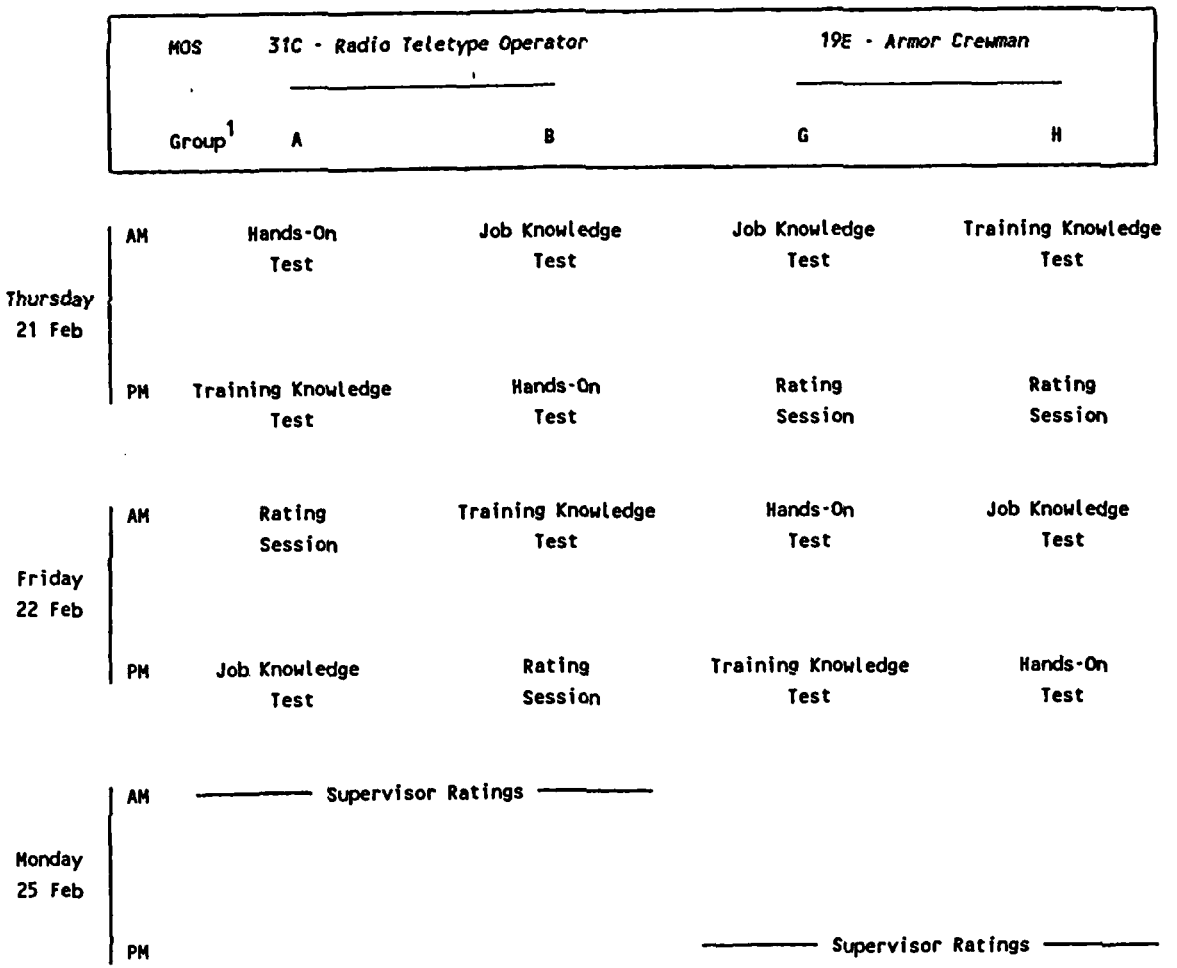
Sample

Before scheduling the field test sites, we constructed a roster of possible first-term enlistees for each MOS. This roster was generated by identifying soldiers whose enlistment date fell between 1 April 1982 and 30 June 1983. This period was selected so that soldiers participating in the field tests would have from fifteen months up to three years of experience on the job. For each field test site, we generated a list of soldiers for each MOS whose entry date fell within this period. (This information was obtained from the World Wide Personnel Locator Service compiled by the U.S. Army.) This list was given to the point-of-contact (POC) at each field test site, who was then responsible for contacting the appropriate units and obtaining the designated number of soldiers from the target MOS on the scheduled days.



[1]Each group equals 15 soldiers in the same MOS.

Figure 3. Field Test Schedule for USAREUR Team 2



[1]Each group equals 15 soldiers in the same MOS.

Figure 4. Field Test Schedule for Fort Stewart

Our goal for Batch A MOS was to include about 150 soldiers from each MOS in the field test sessions. For Batch B, we attempted to include about 180 soldiers from each MOS.

Table 17 and Table 18 provide descriptive information for Batch A and Batch B MOS soldiers participating in the field test sessions. A breakdown of each MOS sample by location, gender, race, pay grade, and age is provided.

Across the nine MOS, note that for gender, three MOS samples contain no females. Recall that 13B, 11B, and 19E are combat arms MOS, and therefore females are not included. Two MOS, 71L and 91A, contain a fairly high percentage of females (50.0% and 37.7% respectively). The remaining MOS samples contain a much smaller proportion of females (64C--7.1%; 95B--2.6%; 31C--12.8%; and 63B--6.5%).

The method for obtaining information about soldiers' race or ethnic group varied from Batch A to Batch B. As is evident from Tables 17 and 18, participants from Batch A MOS were asked to indicate race by checking (1) white, (2) black, (3) Asian, (4) American Indian, or (5) other. On Table 17, we combined the numbers for Asian and American Indian with the "other" category because there were so few in those categories. For the Batch B field test, we revised the category system. Participants were asked to indicate race or ethnic group membership using the following categories; (1) white; (2) black; (3) Hispanic; and (4) other.

Across the nine MOS, the racial membership of our sample varies greatly. The percentage of whites within each MOS ranges from 50.0 to 91.2 percent. For blacks, the percentage ranges from 5.3 to 42.0 percent. For the "other" category, the percentages range from 0.7 to 7.3 percent. Across the five MOS in Batch B, the percentage of Hispanics ranges from 2.0 to 4.1 percent.

Mean age values for Batch A MOS samples range from 21.4 to 22.4 with a median value of 21 for three MOS and 22 for one MOS. The modal age is 20. For Batch B samples the mean age ranges from 22.3 to 23.1, with a median value of 22 for all five MOS. The modal age for these MOS is 21. Since the Batch B field test sessions were conducted six months after the Batch A sessions, we would expect Batch B MOS samples to be slightly older than Batch A MOS samples.

Across the nine MOS, the majority of participants indicated that their pay grade at the time of testing was either E-3 or E-4. The percentage of soldiers in the E-3 and E-4 pay grades ranges from 86.1 percent for Military Police (95B) to 95.5 percent for Motor Transport Operator (64C). A smaller percentage reported pay grades of E-1 or E-2, in only one MOS, Military Police (95B), does the total percentage for these pay grades exceed 10%. Finally, a much smaller percentage of soldiers reported pay grades of E-5 (2.5% for Armor Crewman and 1.4% for Radio Teletype Operator).

The final variable, location, indicates the number of soldiers participating at each field test site. In Batch A, soldiers in the Cannon Crewman (13B) and the Motor Transport Operator (64C) positions were obtained exclusively from OCONUS (USAREUR) locations. Administrative Spe-

Table 17

Description of Field Test Sample by MOS - Batch A

		MOS			
		13B	64C	71L	95B
TOTAL	N	150	155	129	114
GENDER	N	150	155	129	114
Female	N	0	11	64	3
	%	0%	7.1%	50.0%	2.6%
Male	N	150	144	65	111
	%	100%	92.9%	50.0%	97.4%
RACE	N	150	155	129	114
Black	N	63	30	60	6
	%	42%	19.4%	46.5%	5.3%
White	N	84	117	64	104
	%	56%	75.5%	50.0%	91.2%
Other	N	3	8	5	4
	%	2%	5.2%	3.9%	3.5%
AGE	N	150	155	129	114
Mean		21.6	22.4	22.2	21.4
Median		21.0	22.0	21.0	21.0
Mode		20.0	20.0	20.0	20.0
S.D.		2.17	2.74	2.86	2.26
Range		19 - 33	19 - 36	19 - 35	19 - 32
PAY GRADE	N	150	155	67[a]	72[a]
E1	N	4	1	1	1
	%	2.7%	0.6%	1.5%	1.4%
E2	N	7	6	6	9
	%	4.7%	3.9%	9.0%	12.5%
E3	N	62	22	20	46
	%	41.3%	14.2%	29.9%	63.9%
E4	N	77	126	40	16
	%	51.3%	81.3%	59.7%	22.2%
LOCATION	N	150	155	129	114
Fort Hood	N	0	0	48	42
	%	0%	0%	37.2%	36.8%
Fort Polk	N	0	0	60	42
	%	0%	0%	46.5%	36.8%
Fort Riley	N	0	0	21	30
	%	0%	0%	16.3%	26.3%
USAREUR	N	150	155	0	0
	%	100%	100%	0%	0%

[a]We have Pay Grade information for only a subset of the 71L and 95B samples.

Table 18

Description of Field Test Sample by MOS - Batch B

		MOS				
		11B	19E	31C	63B	91A
TOTAL	N	178	172	148	156	167
GENDER	N	178	172	148	153	167
Female	N	0	0	19	10	63
	%	0%	0%	12.8%	6.5%	37.7%
Male	N	178	172	129	143	104
	%	100%	100%	87.2%	93.5%	62.3%
RACE	N	178	172	148	156	167
Black	N	57	36	53	36	48
	%	32.0%	20.9%	35.8%	23.1%	28.7%
Hispanic	N	5	7	3	4	4
	%	2.8%	4.1%	2.0%	2.6%	2.4%
White	N	103	124	91	111	106
	%	57.9%	72.1%	61.5%	71.2%	63.5%
Other	N	13	5	1	5	9
	%	7.3%	2.9%	0.7%	3.2%	5.4%
AGE	N	169	164	139	155	152
Mean		22.4	22.5	22.5	22.3	23.1
Median		22.0	22.0	22.0	22.0	22.0
Mode		21.0	21.0	21.0	21.0	21.0
S.D.		2.72	2.22	2.35	2.79	3.05
Range		19 - 32	19 - 33	18 - 38	19 - 38	18 - 34
PAY GRADE	N	171	162	140	154	151
E1	N	3	1	2	4	1
	%	1.8%	0.6%	1.4%	2.6%	0.7%
E2	N	8	4	7	11	13
	%	4.7%	2.5%	5.0%	7.1%	8.6%
E3	N	33	32	31	38	27
	%	19.3%	19.8%	22.1%	24.7%	17.9%
E4	N	127	121	98	101	110
	%	74.2%	74.7%	70.0%	65.6%	72.8%
E5	N	0	4	2	0	0
	%	0%	2.5%	1.4%	0%	0%
LOCATION	N	178	172	148	156	167
Fort Lewis	N	29	30	16	13	24
	%	16.3%	17.4%	10.8%	8.4%	14.4%
Fort Polk	N	30	31	26	26	30
	%	16.9%	18.0%	17.6%	16.7%	18.0%
Fort Riley	N	30	24	26	29	34
	%	16.9%	14.0%	17.6%	18.6%	20.4%
Fort Stewart	N	31	30	23	27	21
	%	17.4%	17.4%	15.5%	17.3%	12.6%
USAREUR	N	58	57	57	61	58
	%	32.6%	33.1%	38.5%	39.1%	34.7%

cialist (71L) and Military Police (95B) samples were tested exclusively in CONUS locations. Batch B MOS samples were obtained from both CONUS and OCONUS locations.

Preparation for Rating Sessions

Our plan for administering performance ratings included obtaining evaluations from first-term enlistees' colleagues or peers and from enlistees' supervisors. Procedures for identifying an enlistee's peers and supervisors are described below.

Identifying Peers. On Day One of the field test session, we convened the entire group of 60 first-term enlistees to describe the purpose of Project A, the activities they would be involved in over the two day period, and how those activities meshed with the goals of Project A.

Also at this time, the soldiers were given an alphabetized list of recruits from their MOS who were participating in the field test session. They were asked to review the list and to identify as many soldiers as they could whom they had worked with or knew well enough to rate in several job performance areas. We defined a work colleague or peer as: (1) someone they had known for at least two months, and (2) someone they had observed performing on the job on several occasions.

Soldiers were first asked to find their own name on the list and circle it. Next, they were asked to identify the soldiers that they knew by placing a check next to each soldier's name. We asked them to check off as many names on the list as they could, but we also informed them that they would only be asked to rate, at most, four of their peers, regardless of the number they reported knowing.

We used the information on these lists to make peer rating assignments. For the most part, peer assignments were made via computer. A computer program was developed to randomly assign ratees to raters using the information soldiers gave us about individuals with whom they had worked on the job. To operate this program, we first input the information from each soldier's list indicating all enlistees he/she reported knowing well enough to evaluate. The computer program used this information to assign ratees to raters. The output, all things being equal, assigned each rater four ratees or soldiers and assigned each ratee or soldier to four raters. The goal was to obtain four peer ratings for each soldier participating in the field test session.

This procedure required about one-and-one-half hours to complete. After the computer generated the rating assignments, we recorded the names of the ratees on a rating tab along with the name of the rater. Because so much time was required to perform these rating assignments, no rating sessions were conducted during the morning session of Day One.

Identifying Supervisors. First-term enlistees' supervisors were identified by the POC or other military personnel located at each site or post. Our goal was to obtain at least two supervisory ratings for each enlistee

attending the field test sessions. We asked units from which the first-term enlistees were selected to identify the NCO directly responsible for supervising each enlistee as well as the NCO or officer serving as the second-line supervisor for each enlistee.

Thus, when we tested 60 soldiers from an MOS at a particular post, it was possible to have as many as 120 supervisors scheduled to evaluate their performance. In most cases, however, supervisors were able to rate several soldiers. Supervisor rating sessions were conducted with groups of varying sizes, ranging from as few as five to as many as 30 supervisors.

Procedures for Administering Rating Scales

Procedures followed for the peer rating sessions and for the supervisor rating sessions were virtually identical. During each session, participants were asked to evaluate ratees on Army-wide tasks or tasks common to all MOS, Army-wide behaviorally anchored rating scales (BARS) representing broad performance requirements that cut across all MOS, MOS-specific task scales, and MOS-specific BARS. Participants were also asked to complete two questionnaires designed to obtain information about their job history and current job situation. (Documentation of Army-wide rating scale development activities has been prepared by Pulakos & Borman, 1986. Campbell et al., 1986, have documented information for the MOS-specific task rating scales. Olson & Borman, 1986, document the development and results for the Army environment questionnaire.) Below we describe the general procedures for administering these rating scales.

Rating Session. Administrators began each rating session with a brief review of Project A and a description of the activities involved in the rating session. Participants were again reminded that the information they provided would remain strictly confidential and would not appear in their permanent record, nor would anyone in the Army ever be informed of how they had rated their peers or how their peers had evaluated them. Supervisors were informed that their subordinates would never see the ratings they provided and that the ratings would not appear in the enlistees' permanent files.

Next, we gave each participant a rating tab listing the peers or subordinates they would be rating. We asked them to review the list to make sure that they felt confident rating the job performance of all persons on their list. Participants were reminded that we wanted them to only rate soldiers whom they: (1) had known for at least two months and (2) had observed performing on the job. Administrators consulted with each participant who reported problems and resolved these by finding a replacement ratee or by simply deleting a ratee if no replacements were available.

Administrators then distributed the first rating scale booklet. Before participants began making their ratings, administrators provided guidance and instruction about evaluating job performance.

Rater Training. Administrators began this part of the rating session by describing the steps followed in developing the rating scales. They informed participants that the behaviorally anchored rating scales had been

developed with the help of NCOs familiar with the job or MOS in question. That is, the performance dimensions and anchors had been defined by individuals most familiar with MOS job requirements. Next, administrators explained how to use the information provided in the booklets to make their ratings. This included a discussion of the behavioral anchors and an example of how a rater should use these anchors to evaluate ratees' performance.

Finally, administrators discussed four common rating errors and ways to avoid them when providing performance ratings. These errors included: (1) halo error, or failing to consider a person's strengths and weaknesses independently for each performance dimension; (2) single-time error, or basing one's ratings for a person on a single event, failing to consider performance on several occasions; (3) stereotype error, or providing performance ratings based on appearance, background, or other characteristics unrelated to job performance; and (4) same-level-of-effectiveness error, or failing to distinguish between two or more ratees on a single performance dimension.

During this discussion, administrators defined each type of error and provided a relevant example of how it might occur. They emphasized that participants should rely on their observations of each ratee and avoid considering other unrelated factors. Participants were encouraged to ask questions about rating procedures and to obtain clarification on how to avoid the common rating errors.

At the end of this discussion, administrators explained the procedures for recording ratings in the booklets and indicated that they would review the ratings as participants progressed through the booklet answering any questions and dealing with any problems that might arise.

We had three objectives for the rater training session. First, we wanted to ensure that all participants understood the instructions and knew how to record their ratings in the booklet. Second, we wanted to make sure that participants understood the rationale behind the behaviorally anchored rating scales, so that all raters would be using the same "frame of reference" or standards to evaluate ratees' performance. And third, we wanted to ensure that raters understood the importance of reading performance dimension definitions and anchors, and carefully considering the job performance behaviors they had observed, BEFORE evaluating ratees' performance.

We explored the effects of different types of training during the field test sessions. Information about the different types of rater training programs and their impact on peer and supervisor ratings are presented in Pulakos and Borman (1986) and Pulakos (1986).

Administering the Remaining Scales. For the other rating scales included in the workshops, administrators followed essentially the same procedures. They described how the scales had been developed and the procedures for recording ratings on the form or in the booklet provided. Further, raters were reminded that they should try to avoid making the common rating errors, and that because the ratings were for research purposes only, they should be as candid as possible in making their ratings.

Data Analyses

Computing Rating Scores. Ratings collected during the field test sessions were pooled across locations for each MOS. For example, ratings collected for the Armor Crewman position at the five test sites--Fort Lewis, Fort Polk, Fort Riley, Fort Stewart, and USAREUR--were combined and analyzed as a single unit.

One apparent problem with the ratings surfaced when we compared mean ratings for a single ratee provided by two or more raters. Although raters appeared to agree on a particular ratee's strengths and weaknesses across the different performance dimensions, level differences in mean ratings appeared. Because we were more interested in an enlistee's profile of ratings across the different performance dimensions (i.e., a ratee's relative strengths and weaknesses), we decided to compute adjusted scores that would reduce or eliminate the level differences between scores provided by two or more raters for a single ratee.

An examination of the ratings provided by each rater revealed that some raters had failed to provide ratings for all enlistees on each performance dimension. Therefore, it was necessary to compute adjusted scores by comparing raters' evaluations on a single performance dimension rather than across all performance dimensions. Below we describe the procedures developed to compute adjusted ratings or scores; we include an example for one rater and one performance dimension to demonstrate how these adjustments were made.

- For each rater, we identified the score provided for one enlistee on a single performance dimension. For example, Rater 1 gave Enlistee A a score of 4.0 and Enlistee B a score of 5.0 on Dimension X.

- We identified all other peer and supervisor raters providing evaluations for the same enlistees on that same performance dimension as the target rater. For each enlistee, we computed the mean rating across all raters. In our example, Raters 2, 3, and 4 evaluated enlistee A on Dimension X; we computed the mean rating for enlistee A across these three raters, for a mean of 5.3. Only two raters, Raters 3 and 4, evaluated Enlistee B on Dimension X; we calculated the mean rating for Raters 3 and 4 for Enlistee B; for a mean of 5.5.

- We then compared the score for the target rater-enlistee pair with the mean computed for the same enlistee across all other raters. These values were used to compute a mean difference score for the target rater-enlistee pair. Continuing with our example, Rater 1 gave Enlistee A a rating of 4.0 while the other three raters evaluating Enlistee A provided a mean rating of 5.3. Thus Rater 1 would receive a difference score of -1.3 for Enlistee A on Dimension X.

- This procedure was repeated to compute a difference score for each rater-enlistee combination on each performance dimension. Values for Enlistee B are 5.0 for Rater 1 and 5.5 for Raters 3 and 4, giving Rater 1 a mean difference score of -0.5 for Enlistee B on Dimension X.

- For each target rater-enlistee pair, we identified a value for weighting the difference score. In our example, Rater 1 has a difference score of -1.3 for Enlistee A and -0.5 for Enlistee B. We weighted each score using the number of other raters evaluating each enlistee. So, in this example the mean difference score for Enlistee A is weighted 3 because three other raters evaluated this enlistee. The mean difference score for Enlistee B is weighted 2.

- For each rater, we computed a weighted average difference score for each performance dimension. For Dimension X, Rater 1 received a weighted average difference score of -1.0 [i.e., (3 (-1.3) + 2 (-0.5))/5].

- Finally, an average difference score was computed across all performance dimensions for that rater. The average difference score was then used to adjust all ratings provided by the target rater. For Rater 1 the average across all performance dimensions is -1.2. Therefore, all ratings provided by Rater 1 were increased by a value of 1.2.

The above procedures were used to compute adjusted scores for all raters. Ratings supplied by peers and supervisors were pooled to compute adjusted scores.

Screening the Rating Data. The next step in the analyses involved screening the data to identify ratings that appeared unrealistic or did not correspond to other ratings provided for the same ratee. Because "true" performance scores were not available, we evaluated the data by comparing information provided by one rater with information provided by all other raters evaluating the same enlistee(s). Two criteria for identifying questionable raters were developed.

- First, we computed the correlation between performance dimension ratings for a target rater-enlistee pair and the mean performance dimension ratings provided by all other raters evaluating that enlistee. If this correlation was -.2 or lower *for any enlistee*, *all* of the rater's ratings were deleted from the data set.

- Second, we examined each rater's average difference score used to make the rating score adjustments. Any rater that obtained an average difference score of 2.0 or greater in absolute value was deleted from the sample.

For any rater whose adjusted scores met one or both of the above screening criteria, *all* ratings provided by that rater were deleted from the data set. Thus, for one discrepant rater, we may have eliminated one or more ratees. This number varied according to the number of soldiers evaluated

by the discrepant rater.

Our goal for eliminating raters was to be as conservative as possible by deleting only the most extreme ratings. As a result, very few ratees were deleted from the data set. For each of the MOS by rater type (supervisors or peers) data sets, the number of ratees deleted from set ranges from zero to seven. Across all MOS and rater types, data were eliminated for only 22 ratees.

Subsequent Analyses. For all remaining analyses, we analyzed ratings provided by supervisors separately from ratings provided by peers. Using the adjusted scores computed for each rater, we computed a mean performance dimension score for each ratee. These mean values were used to compute the mean, standard deviation, and range of scores across all ratees for each performance dimension.

We computed the intraclass correlation between ratings provided for the same enlistees to estimate the degree of interrater reliability on each performance dimension. Next, intercorrelations between performance dimension ratings provided by peers and between performance dimension ratings provided by supervisors were computed. Intercorrelations between peer and supervisor ratings were also computed. We present and discuss these data separately for each MOS in the "Results" section.

Differences Between Batch A and Batch B Data Sets. Before presenting these data, however, we must call attention to some differences between the adjusted rating scores computed for Batch A MOS and Batch B MOS.

First, recall that for all MOS, raters used a scale of 1 (low) to 7 (high) to evaluate ratees. These "raw" ratings were then adjusted for level differences between raters, using the procedure described above. This procedure provided some adjusted scores that fell outside the actual range of rating scale values; for example, the rating scores for one performance dimension ranged from 0.49 to 7.17. In the analyses of Batch A MOS ratings, we allowed the adjusted values to exceed the actual scale point range. For Batch B MOS, we modified the adjusted scores so that the range of adjusted values would correspond to the range of "raw" values (i.e., all scores would fall within a range of 1 to 7); this was accomplished by truncating adjusted scores that exceeded 7.0 or that fell below 1.0. In the following tables, then, the ratings for Batch A exceed the range of 1 to 7, whereas ratings for Batch B MOS fall within this range.

Another difference in the analyses performed for the two batches of MOS involves the assumptions made in computing the interrater reliability estimates for peers. Since the goal was to obtain four peer ratings for each enlistee, in computing the interrater reliability coefficients for peer ratings obtained for Batch A MOS we assumed four raters per ratee. When computing these values for Batch B MOS, we first computed the average number of peer raters per ratee. This information led us to modify our assumption about the average number of raters, so for Batch B MOS interrater reliability estimates were computed assuming three raters per ratee.

Interrater reliability estimates computed for peer ratings provided for Batch A MOS samples can be interpreted as the expected correlation between

(1) the mean ratings provided for soldiers by their peers in this sample and (2) the mean ratings that would be provided for the same soldiers by an equivalent group of peers, assuming that all soldiers were rated by four peers. "Equivalent" indicates any peer who meets the two criteria for rating a soldier.

Interpretation of interrater reliability estimates computed for peer ratings provided for Batch B MOS samples is similar to the interpretation for Batch A MOS, except that we assume that three rather than four peers provided ratings for Batch B.

For *all* MOS, interrater reliability estimates computed for supervisors can be interpreted as the expected correlation between (1) the mean ratings provided for soldiers by their supervisors in this sample and (2) the mean ratings that would be provided for the same soldiers by an equivalent group of supervisors, assuming that all soldiers were rated by two supervisors. By "equivalent," we mean any supervisor who meets the two criteria for rating a soldier.

Assumptions concerning the number of raters evaluating each soldier affect the resulting reliability estimate. The more raters evaluating a soldier, generally, the higher the estimate. For the field test data, then, we would expect higher interrater reliability estimates for ratings provided by peers than by supervisors, and higher reliability estimates for ratings provided by peers in Batch A MOS than by peers in Batch B MOS.

Results

For each group of ratings, we had calculated the ratio of the number of raters to the number of ratees. These data, reported in Table 19, are presented separately for each MOS and for supervisor and peer ratings. For comparison, we have included ratios for rating data computed before and after the ratings were screened. Note that these ratios change very little following the screening process.

For supervisors, the "after" ratios range from 1.04 for Administrative Specialist (71L) to 1.88 for Military Police (95B) with a median value of 1.73. These data indicate that for a majority of enlistees in each MOS, we obtained ratings from two supervisors. Within the Administrative Specialist MOS, however, we obtained an average of only one supervisor rating for each enlistee.

For peer ratings, the "after" ratio of raters to ratees ranges from 1.89 for Administrative Specialist (71L) to 3.39 for Military Police (95B) with a median value of 2.57. Thus, we obtained at least two peer ratings for every enlistee with the exception of Administrative Specialist enlistees. For enlistees in four of the MOS, Military Police (95B), Infantryman (11B), Armor Crewman (19E), and Medical Specialists (91A), we obtained about three peer ratings for each.

On the following pages, we describe the results for each MOS individually. For each rater group (i.e., supervisors and peers), we report the range of adjusted ratings, mean, and standard deviation for each performance dimen-

Table 19

Ratio of Raters to Ratees Before and After Screening for Supervisor and Peer Ratings

MOS	Supervisors		Peers	
	Before	After	Before	After
13B - Cannon Crewman	1.47	1.47	2.89	2.52
64C - Motor Transport Operator	1.84	1.82	2.77	2.57
71L - Administrative Specialist	1.04	1.04	1.90	1.89
95B - Military Police	1.94	1.88	3.67	3.39
11B - Infantryman	1.81	1.81	2.99	2.99
19E - Armor Crewman	1.68	1.68	2.95	2.95
31C - Radio Teletype Operator	1.73	1.73	2.49	2.50
63B - Light-Wheel Vehicle Mechanic	1.77	1.77	2.08	2.09
91A - Medical Specialist	1.59	1.59	3.10	3.10

sion as well as the grand mean across all performance dimensions and ratees. For comparison, the text includes the grand mean computed across unadjusted ratings (this value does not appear in the tables). We also focus on the interrater reliability estimates (R_{xx}) and the intercorrelations between performance dimension ratings provided by peers and by supervisors.

Cannon Crewman - 13B

We collected performance information on a total of 150 first-term enlistees from the Cannon Crewman MOS. Table 20 presents the means, standard deviations, range of scores and interrater reliability estimates for supervisors and peers.

Complete supervisor rating data were collected for 140 enlistees. Focusing on those ratings, adjusted ratings range from 0.65 to 7.76. Mean adjusted performance dimension values range from 4.48 to 5.19 (standard deviations range from 1.03 to 1.31). The grand mean, computed across all enlistees and all performance dimensions, using the adjusted ratings, is 4.89 (SD=0.81). The unadjusted grand mean value is 4.89 (SD=1.13). Interrater reliability estimates range from .33 (J. Position improvement) to .70 (K. Overall performance) with a median value of .45.

Ratings provided by peers, for 140 enlistees, adjusted for level differences, range from 0.76 to 8.87. Mean adjusted ratings across the 11 performance dimensions range from 4.47 to 5.05 and the standard deviations range from 0.80 to 1.22. The grand mean value computed for adjusted scores is 4.85 (SD=0.71); the grand mean for unadjusted values is 4.89 (SD=0.84). Reliability estimates range from .40 (H. Receiving and relaying communications) to .66 (G. Loading/unloading Howitzer) with a median value of .54.

Table 21 presents the intercorrelation matrix for the supervisor and peer ratings. For supervisors alone, correlations between the dimension ratings (excluding Overall performance) range from .19 to .70 with a mean value of .46 (SD=0.12). Examination of the Overall ratings provided by supervisors indicates that "Gunnery" (Dimension F), "Position improvement" (Dimension J), and "Loading/unloading Howitzer" (Dimension G) correlate highest with this rating.

Correlations between dimension ratings provided by peers (excluding Overall performance) range from .36 to .62 with a mean of .50 (SD=0.07). For peers, "Gunnery" (Dimension F), "Recording/ record keeping" (Dimension I), and "Position improvement" (Dimension J) correlate highest with the Overall rating.

Intercorrelations between dimension ratings provided by supervisors and by peers (excluding Overall performance) range from .15 to .53. The degree of agreement between peers and supervisors is more apparent from the values in the diagonal of this matrix (e.g., peer ratings on Dimension A correlated with supervisor ratings on Dimension A). Correlations between supervisor and peer ratings on the 11 performance dimensions range from .18 (E. Setting Up Communications) to .54 (D. Preparing for occupation/emplacing Howitzer) with a median value of .39.

Table 20

Means, Standard Deviations, Ranges, and Reliability Estimates for

Cannon Crewman (13B) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Loading Out Equipment	141	5.02	1.17	0.65 - 7.67	.58	140	4.83	0.85	2.19 - 6.63	.54
B. Driving and Maintaining Vehicles, Howitzers and Equipment	141	5.14	1.17	1.67 - 7.67	.63	141	5.05	1.00	1.76 - 7.06	.59
C. Transporting/Sorting/Storing and Preparing Ammunition for Fire	141	4.98	1.09	0.65 - 6.86	.61	141	5.04	0.90	2.28 - 6.69	.59
D. Preparing for Occupation/ Emplacing Howitzer	141	4.89	1.14	1.48 - 6.87	.45	141	4.71	0.98	2.21 - 6.87	.53
E. Setting Up Communications	141	4.81	1.04	2.48 - 7.67	.43	141	4.95	0.80	2.19 - 6.69	.42
F. Gunnery	141	4.48	1.26	1.42 - 7.76	.44	141	4.47	1.22	0.76 - 7.87	.54
G. Loading/Unloading Howitzer	141	5.19	1.10	0.65 - 7.67	.35	141	5.02	1.05	2.05 - 8.87	.66
H. Receiving and Relaying Communications	141	4.84	1.11	1.48 - 6.87	.40	141	4.94	0.83	2.05 - 6.87	.40
I. Recording/Record Keeping	140	4.64	1.31	0.65 - 7.36	.61	141	4.61	1.03	2.05 - 6.87	.61
J. Position Improvement	141	4.95	1.03	1.48 - 7.67	.33	141	4.92	0.93	2.19 - 7.87	.51
K. Overall Cannon Crewman Performance	141	4.92	1.15	0.65 - 7.67	.70	141	4.91	0.95	2.19 - 7.87	.61
Mean Ratings	141	4.89	0.81	1.88 - 6.81	.73	141	4.85	0.71	2.49 - 6.97	.80

Table 21

Supervisor and Peer Intercorrelations for

Cannon Crewman (13B) MOS-Specific BARS

		Supervisors											Peers									
		A	B	C	D	E	F	G	H	I	J	K	A	B	C	D	E	F	G	H	I	J	K
S	A. Load Equipment																					
U	B. Drive/Maintain Equipment	61																				
P	C. Transport/Sort/Store Ammo	69	62																			
E	D. Prepare/Emplace Howitzer	46	44	59																		
R	E. Set Up Communications	50	58	48	41																	
V	F. Gunnery	43	40	33	20	46																
I	G. Load/Unload Howitzer	55	36	51	41	42	66															
S	H. Receive/Relay Communications	41	36	53	70	32	30	43														
O	I. Record/Record Keeping	40	23	46	54	19	42	39	60													
R	J. Position Improvement	44	44	58	44	56	46	48	47	40												
S	K. Overall	53	43	49	31	54	66	61	30	35	65											
	A. Load Equipment	50	36	47	53	28	31	42	47	31	35	38										
	B. Drive/Maintain Equipment	46	39	45	35	24	31	44	28	26	30	34	62									
	C. Transport/Sort/Store Ammo	48	29	46	42	31	24	45	29	32	37	34	52	58								
P	D. Prepare/Emplace Howitzer	27	20	26	54	27	15	31	48	34	31	27	42	36	40							
E	E. Set Up Communications	30	16	16	31	18	15	21	27	25	19	22	51	39	44	60						
E	F. Gunnery	36	28	37	45	22	38	39	34	43	37	37	45	51	52	53	46					
R	G. Load/Unload Howitzer	27	24	31	46	25	21	30	51	27	29	18	52	37	51	61	53	51				
S	H. Receive/Relay Communications	34	24	24	44	24	23	31	41	22	21	28	54	49	45	58	62	47	61			
	I. Record/Record Keeping	35	22	21	41	25	30	33	37	33	18	28	55	40	43	55	59	43	46	60		
	J. Position Improvement	40	35	42	47	36	25	42	39	33	37	39	48	51	55	50	46	53	45	49	44	
	K. Overall	42	32	39	46	35	41	39	42	41	42	44	62	57	48	56	55	68	58	62	63	63
		A	B	C	D	E	F	G	H	I	J	K	A	B	C	D	E	F	G	H	I	J	K

Motor Transport Operator - 64C

A total of 155 enlistees from the Motor Transport Operator position participated in the field test sessions. Means, standard deviations, range of scores and interrater reliability estimates are presented in Table 22 for supervisor and peer ratings.

We gathered supervisor ratings on all performance dimensions for 138 of these enlistees. Across all dimensions supervisor ratings adjusted for level differences, range from 0.49 to 7.94. Mean adjusted scores range from 4.16 to 5.11 (standard deviations range from 0.92 to 1.12). The grand mean computed across all enlistees and performance dimensions, for the adjusted ratings, is 5.07 (SD=0.73); the grand mean computed for unadjusted ratings is 4.92 (SD=1.02). Interrater reliability estimates range from .47 (F. Parking and securing vehicles) to .66 (I. Safety-mindedness and E. Loading cargo and transporting personnel) with a median value of .57.

The peer rating data indicate that we obtained complete data for 152 enlistees. Adjusted scores range from 0.17 to 8.49. Mean adjusted ratings for individual performance dimensions range from 3.78 to 5.39 (standard deviations range from 0.75 to 1.09). The grand mean computed for adjusted ratings provided for all enlistees across all performance dimensions is 4.74 (SD=0.66); for unadjusted ratings the grand mean is 4.66 (SD=0.83). Interrater reliability estimates range from .32 (G. Performing administrative duties) to .68 (D. Using maps/following proper routes) with a median value of .54.

The supervisor and peer intercorrelation matrix appears in Table 23. Correlations computed for supervisor ratings alone (excluding Overall performance) range from .21 to .65 with a mean of .48 (SD=0.12). Correlations between the final dimension, "Overall," and the other performance dimensions indicate that supervisors placed the highest value on "Loading cargo and transporting personnel" (Dimension E), "Safety-mindedness" (Dimension I), and "Checking and maintaining vehicles" (Dimension C).

Correlations computed between performance dimension ratings provided by peers (excluding Overall performance) range from .09 to .69 with a mean of .42 (SD=0.16). For the peer group, "Driving vehicles" (Dimension A), "Safety-mindedness" (Dimension I), and "Checking and maintaining vehicles" (Dimension C) correlate highest with the Overall performance rating.

Intercorrelations between supervisor and peer ratings (excluding the Overall rating) range from .06 to .54. The level of agreement between supervisor and peer ratings is apparent from the 11 correlations highlighted in the diagonal of the matrix. These values range from .20 (J. Performing dispatcher duties) to .53 (C. Checking and maintaining vehicles) with a median of .46.

Table 22

Means, Standard Deviations, Ranges, and Reliability Estimates for

Motor Transport Operator (64C) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Driving Vehicles	147	5.06	1.01	1.39 - 6.97	.54	154	4.92	1.08	0.17 - 8.49	.63
B. Vehicle Coupling	149	4.80	1.12	1.39 - 7.17	.57	154	4.67	0.99	1.17 - 7.49	.53
C. Checking and Maintaining Vehicles	149	4.91	1.12	0.49 - 6.97	.48	154	4.98	0.94	2.03 - 7.19	.58
D. Using Maps/Following Proper Routes	147	4.84	0.97	0.94 - 7.08	.59	154	4.69	1.05	1.76 - 6.91	.68
E. Loading Cargo and Transporting Personnel	147	4.91	0.98	1.32 - 7.02	.66	153	4.88	0.88	2.03 - 7.49	.54
F. Parking and Securing Vehicles	147	5.11	1.04	1.39 - 7.08	.47	154	5.39	0.85	2.49 - 7.69	.42
G. Performing Administrative Duties	147	4.91	0.97	1.89 - 7.94	.56	154	4.87	0.75	2.49 - 7.19	.32
H. Self-Recovering Vehicles	145	4.40	1.02	1.73 - 6.39	.53	154	4.23	0.93	1.13 - 6.32	.54
I. Safety-Mindedness	148	4.85	1.00	1.88 - 6.49	.66	154	5.00	0.92	2.01 - 7.49	.60
J. Performing Dispatcher Duties	138	4.16	1.10	1.52 - 6.47	.60	152	3.78	1.09	0.59 - 6.23	.36
K. Overall Motor Transport Operator Performance	147	4.82	0.92	1.82 - 6.60	.62	154	5.03	0.95	1.61 - 7.49	.58
Mean Ratings	154	5.07	0.73	2.80 - 6.70	.74	154	4.74	0.66	2.49 - 6.79	.82

Table 23

Supervisor and Peer Intercorrelations for

Motor Transport Operator (64C) MOS-Specific BARS

		Supervisors											Peers									
		A	B	C	D	E	F	G	H	I	J	K	A	B	C	D	E	F	G	H	I	J	K
S	A. Drive Vehicles																					
U	B. Vehicle Coupling	51																				
P	C. Check/Maintain Vehicles	61	59																			
E	D. Use Maps/Follow Proper Routes	62	46	46																		
R	E. Load Cargo/Transport Personnel	59	63	64	61																	
V	F. Park/Secure Vehicles	39	53	65	36	57																
I	G. Perform Administrative Duties	46	33	56	41	57	58															
S	H. Self-Recover Vehicles	41	41	53	48	46	45	45														
O	I. Safety-Mindedness	59	46	59	40	59	64	64	47													
R	J. Perform Dispatcher Duties	26	23	33	21	22	31	50	47	45												
S	K. Overall	61	61	71	53	73	63	67	48	73	42											
	A. Drive Vehicles	49	47	53	39	40	48	38	39	45	35	46										
	B. Vehicle Coupling	42	52	46	26	40	48	36	23	42	18	47	56									
	C. Check/Maintain Vehicles	49	40	53	37	47	49	45	30	45	27	49	64	55								
P	D. Use Maps/Follow Proper Routes	42	37	44	46	45	40	33	42	35	23	42	51	34	51							
E	E. Load Cargo/Transport Personnel	47	47	54	42	49	46	30	38	36	30	45	69	51	64	54						
E	F. Park/Secure Vehicles	47	49	47	39	48	34	35	33	37	24	36	63	48	45	49	60					
R	G. Perform Administrative Duties	35	35	41	34	42	43	43	35	41	33	45	27	33	41	48	38	30				
S	H. Self-Recover Vehicles	46	39	46	37	35	30	33	41	38	30	48	35	36	51	46	42	24	48			
	I. Safety-Mindedness	40	38	45	36	30	35	44	27	40	27	42	63	48	59	52	52	58	37	43		
	J. Perform Dispatcher Duties	17	30	29	6	27	16	22	11	26	20	26	9	18	11	18	17	20	36	28	17	
	K. Overall	55	45	59	45	44	47	46	34	50	33	51	71	53	66	57	64	56	40	59	67	26
		A	B	C	D	E	F	G	H	I	J	K	A	B	C	D	E	F	G	H	I	J	K

Administrative Specialist - 71L

A total of 129 first-termers from the Administrative Specialist MOS participated in the field test. Table 24 contains performance dimension means, standard deviations, range of scores, and interrater reliability estimates for ratings provided by supervisors and peers.

Results from the supervisor ratings indicate that we obtained complete data for only 95 enlistees. This information suggests the unique circumstances surrounding this MOS. First, enlistees in this MOS often work alone with only one NCO, officer, or civilian providing daily or routine supervision; it was difficult to locate two supervisors for each enlistee. Second, enlistees performing as Administrative Specialists perform some but not all duties delegated to this MOS; thus, raters simply could not rate enlistees on all dimensions. For this MOS, then, we generally obtained enlistee performance ratings from only one supervisor. Only on rare occasions were we able to obtain two such ratings for an enlistee. (Table 19 indicates that the ratio of raters to ratees is 1.04.) Therefore, we did not calculate interrater reliability estimates for supervisor data.

Results from Table 24 indicate that values for supervisor ratings ranged from 1.00 to 8.03. Mean adjusted scores range from 4.11 to 5.26 (standard deviations range from 1.13 to 1.44). The grand mean computed across all enlistees and performance dimensions, using adjusted ratings, is 4.52 (SD=0.94); the grand mean for unadjusted ratings is 4.56 (SD=1.13).

Data for peer ratings indicate that we had similar problems obtaining complete rating data, because soldiers in this MOS seldom work closely with peers. Thus, we obtained complete data for only 63 enlistees but we did collect a sufficient number of ratings to estimate reliabilities for peer rating data. (Table 19 indicates that we obtained 1.89 peer ratings for each enlistee.)

Adjusted peer ratings range from 1.56 to 7.31. Mean adjusted performance dimension ratings range from 4.32 to 5.48 (standard deviations range from 0.81 to 1.09). The grand mean computed across all enlistees and performance dimensions, using adjusted ratings, is 4.72 (SD=0.64); the grand mean computed for unadjusted ratings is 4.75 (SD=0.81). Interrater reliability estimates range from .37 (H. Providing customer service) to .55 (G. Safeguarding and monitoring security of classified documents, and I. Overall performance) with a median value of .49.

The intercorrelation matrix for supervisor and peer ratings is provided in Table 25. For supervisors alone, correlations between the first eight performance dimensions (excluding Overall) range from .15 to .66 with a mean of .42 (SD=0.14). According to the supervisors, "Preparing, typing, and proofreading documents" (Dimension A), "Distributing and dispatching incoming and outgoing documents" (Dimension B), and "Providing customer service" (Dimension H) correlate highest with Overall performance.

For peers alone, correlations between performance dimension ratings (excluding Overall performance) range from .17 to .62 with the mean equal to .36 (SD=0.11). According to the peer ratings, "Providing customer

service" (Dimension H), "Keeping records" (Dimension F), and "Preparing, typing and proofreading documents" (Dimension A) correlate highest with Overall performance.

Intercorrelations between supervisor and peer ratings (excluding correlations with the overall rating) range from .03 to .54. The 10 correlations computed between supervisor and peer ratings on common performance dimensions range from .22 (F. Keeping records, and G. Safeguarding and monitoring security of classified documents) to .51 (I. Overall performance) with a median value of .40.

Table 24

Means, Standard Deviations, Ranges, and Reliability Estimates for

Administrative Specialist (71L) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Preparing, Typing and Proofreading Documents	105	4.48	1.38	1.00 - 7.02	--	64	4.94	0.95	2.50 - 7.22	.50
B. Distributing and Dispatching Incoming and Outgoing Documents	104	4.52	1.23	1.00 - 7.13	--	63	4.49	0.86	2.00 - 6.28	.54
C. Maintaining Office Resources	107	4.66	1.20	2.00 - 7.02	--	63	4.73	0.91	1.96 - 7.31	.44
D. Posting Regulations	105	4.11	1.29	1.00 - 7.02	--	63	4.47	0.88	2.41 - 6.78	.48
E. Establishing and/or Maintaining Files in Accordance with TAFFS	106	4.27	1.44	1.00 - 7.17	--	63	4.55	1.04	1.56 - 7.31	.46
F. Keeping Records	107	4.42	1.25	2.00 - 7.63	--	63	4.77	1.00	2.00 - 6.28	.49
G. Safeguarding and Monitoring Security of Classified Documents	95	4.57	1.13	1.80 - 7.00	--	63	4.32	1.02	1.89 - 6.71	.55
H. Providing Customer Service	107	5.26	1.35	2.00 - 7.63	--	64	5.48	1.09	1.96 - 7.22	.37
I. Overall Administrative Specialist Performance	107	4.85	1.27	1.00 - 8.03	--	64	5.24	0.81	3.00 - 7.00	.55
Mean Ratings	107	4.52	0.94	1.86 - 6.65	--	64	4.72	0.64	2.84 - 5.90	.81

Table 25

Supervisor and Peer Intercorrelations for

Administrative Specialist (71L) MOS-Specific BARS

		Supervisors									Peers								
		A	B	C	D	E	F	G	H	I	A	B	C	D	E	F	G	H	I
S																			
U	A. Prepare/Type/Proofread Documents																		
P	B. Distribute/Dispatch Incoming/Outgoing Documents	61																	
E	C. Maintain Office Resources	46	51																
R	D. Post Regulations	49	59	43															
V	E. Establish/Maintain Files	51	66	48	57														
I	F. Keep Records	52	61	37	41	51													
S	G. Safeguard/Monitor Security of Documents	18	34	15	25	22	31												
O	H. Provide Customer Service	45	42	28	38	38	48	19											
R	I. Overall	69	68	46	62	56	62	30	63										
S																			
	A. Prepare/Type/Proofread Documents	46	43	46	23	40	43	14	29	54									
	B. Distribute/Dispatch Incoming/Outgoing Documents	39	47	29	36	23	32	19	53	42	24								
P	C. Maintain Office Resources	42	40	26	18	27	29	34	26	37	53	33							
E	D. Post Regulations	42	41	30	46	34	3	15	8	31	21	23	34						
E	E. Establish/Maintain Files	41	51	28	23	38	35	25	27	50	39	40	17	32					
R	F. Keep Records	41	48	41	47	50	22	22	38	51	42	50	36	56	44				
S	G. Safeguard/Monitor Security of Documents	33	53	45	39	36	37	22	29	40	24	43	31	23	30	34			
	H. Provide Customer Service	54	46	49	50	40	36	22	40	55	42	36	46	28	19	62	32		
	I. Overall	55	49	51	37	39	46	23	44	51	53	32	47	25	24	54	36	68	
		A	B	C	D	E	F	G	H	I	A	B	C	D	E	F	G	H	I

Military Police - 95B

We tested 114 Military Police enlistees in the field test sessions. Table 26 contains performance dimension rating statistics for supervisor and peer ratings. Note that for both sets of data, we obtained complete data for nearly all subjects (N=111).

Adjusted ratings provided by supervisors range from 1.59 to 7.19. The adjusted means computed for the eight performance dimensions range from 4.12 to 4.77. Adjusted standard deviations for the mean ratings range from 0.82 to 1.03. The grand mean computed using the adjusted ratings is 4.47 (SD= 0.63); for unadjusted ratings the grand mean is 4.59 (SD= 0.75). Interrater reliability estimates range from .39 (B. Providing security) to .74 (H. Overall performance) with a median value of .55.

Peer ratings, adjusted for level differences, range from 1.88 to 7.19. Adjusted mean values computed for each performance dimension range from 4.19 to 4.75 and the standard deviations range from 0.63 to 0.87. The grand mean computed across all enlistees and all performance dimensions, using adjusted ratings, is 4.43 (SD= 0.60); the grand mean computed for unadjusted ratings is 4.43 (SD= 0.66). Interrater reliability estimates range from .39 (B. Providing security) to .71 (H. Overall performance) with a median value of .65.

Table 27 contains the intercorrelations for supervisor and peer ratings. For supervisors alone, these correlations for the seven performance dimensions (excluding Overall) range from .20 to .61 with a mean of .39 (SD= 0.15). According to supervisors, "Investigating crimes/making arrests" (Dimension C), "Providing security" (Dimension B), and "Traffic control and enforcement" (Dimension A) correlate highest with "Overall performance."

Correlations between dimension ratings (excluding Overall) provided by peers range from .48 to .72 with a mean of .58 (SD= 0.07). According to peers, "Traffic control and enforcement" (Dimension A), "Patrolling" (Dimension D), and "Promoting the public image of the Military Police" (Dimension E) correlate highest with Overall performance.

Intercorrelations between supervisor ratings and peer ratings (excluding Overall performance) range from .24 to .54. Correlations computed between peer and supervisor ratings on common performance dimensions range from .31 (G. Responding to medical emergencies) to .55 (H. Overall performance) with a median value of .45.

Table 26

Means, Standard Deviations, Ranges, and Reliability Estimates for

Military Police (95B) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Traffic Control and Enforcement	113	4.67	0.91	2.07 - 6.31	.54	113	4.60	0.69	3.11 - 6.19	.68
B. Providing Security	112	4.77	0.82	2.59 - 6.68	.39	112	4.61	0.64	3.08 - 6.25	.39
C. Investigating Crimes and Making Arrests	113	4.50	0.89	2.59 - 6.98	.57	112	4.42	0.78	2.64 - 6.43	.63
D. Patrolling	111	4.55	1.02	1.59 - 6.67	.56	112	4.51	0.87	1.88 - 7.19	.67
E. Promoting the Public Image of the Military Police	113	4.29	1.03	2.01 - 7.19	.57	112	4.19	0.87	2.28 - 5.88	.66
F. Interpersonal Communications Skills	112	4.36	0.88	1.67 - 7.19	.49	112	4.36	0.81	2.16 - 6.16	.59
G. Responding to Medical Emergencies	112	4.12	0.87	2.17 - 6.09	.52	112	4.38	0.63	2.39 - 5.89	.60
H. Overall Military Police Performance	113	4.57	0.94	1.59 - 6.57	.74	113	4.75	0.76	2.36 - 6.19	.71
Mean Ratings	113	4.47	0.63	2.81 - 5.85	.76	113	4.43	0.60	2.74 - 5.74	.83

Table 27

Supervisor and Peer Intercorrelations for

Military Police (95B) MOS-Specific BARS

		Supervisors								Peers							
		A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
S																	
U	A: Traffic Control/Enforcement																
P	B. Providing Security	58															
E	C. Investigating Crimes/Make Arrests	60	56														
R	D. Patrolling	59	61	53													
V	E. Promoting Public Image	26	24	26	22												
I	F. Interpersonnal Communications Skills	29	36	38	37	48											
S	G. Responding to Medical Emergencies	25	38	46	25	24	20										
O	H. Overall	57	60	69	56	54	52	46									
R																	
S																	
	A. *Traffic Control/Enforcement*	51	51	47	43	34	29	35	53								
P	B. Providing Security	37	46	35	42	32	32	29	43	56							
E	C. Investigating Crimes/Make Arrests	45	44	47	30	42	33	36	52	72	65						
E	D. Patrolling	52	54	53	39	28	26	37	57	66	49	62					
R	E. Promoting Public Immage	42	46	45	37	38	34	40	54	53	63	60	67				
S	F. Interpersonnal Communications Skills	36	43	37	46	24	44	38	47	49	49	50	48	60			
	G. Responding to Medical Emergencies	47	45	36	40	33	29	31	44	61	55	56	51	66	58		
	H. Overall	53	59	52	39	29	32	44	55	75	61	67	74	73	71	68	
		A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H

Infantryman - 11B

A total of 178 enlistees from the Infantryman MOS attended the field test sessions. Table 28 contains the means, standard deviations, range of ratings, and interrater reliability estimates for supervisors and peers. Please note that for this and the remaining MOS, we computed adjusted ratings to remove level differences among raters. These ratings were truncated so that the range of adjusted scores is equivalent to the range of raw or unadjusted scores.

The data in Table 28 indicate that we obtained one or more supervisor ratings for 148 enlistees. Adjusted ratings provided by supervisors range from 1.22 to 7.00. Mean adjusted values computed across all ratees for each performance dimension range from 4.00 to 4.77 (standard deviations range from 0.85 to 1.10). The grand mean computed across all enlistees and performance dimensions for adjusted ratings is 4.45 (SD= 0.70); the grand mean for unadjusted ratings is 4.39 (SD= 0.91). Interrater reliability estimates computed for each performance dimension range from .29 (L. Prisoners of war) to .63 (A. Maintaining supplies, equipment, and weapons) with a median value of .53.

For peer ratings, we obtained complete data for 172 enlistees. Adjusted ratings provided by peers range from 1.76 to 7.00. Mean adjusted values computed across ratees for each performance dimension range from 4.22 to 4.80; standard deviations range from 0.74 to 0.98. The grand mean computed across all enlistees and performance dimensions, using adjusted ratings, is 4.51 (SD= 0.62); the grand mean for unadjusted ratings is 4.56 (SD= 0.70). Interrater reliability estimates range from .30 (G. Avoiding enemy detection) to .64 (C. Navigation) with a median value of .55.

Intercorrelations among supervisor and peer ratings appear in Table 29. For supervisors alone, correlations between dimensions (excluding Overall performance) range from .19 to .65 with a mean of .42 (SD= 0.10). According to the supervisors, "Maintaining supplies, equipment, and weapons" (Dimension A), "Assisting and leading others" (Dimension B), and "Reconnaissance and patrol" (Dimension I) correlate highest with "Overall performance."

For peer ratings alone, correlations for the first 12 dimensions (excluding Overall performance) range from .29 to .63 with a mean value of .50 (SD= 0.08). According to the peer raters, "Use of weapons and other equipment" (Dimension D), "Reconnaissance and patrol" (Dimension I), and "Navigation" (Dimension C) correlate highest with "Overall performance."

Intercorrelations computed between supervisor and peer ratings (excluding Overall performance) range from .11 to .52. Correlations computed for peer and supervisor ratings on common performance dimensions range from .29 (L. Prisoners of war) to .51 (M. Overall performance) with a median value of .41.

Table 28

Means, Standard Deviations, Ranges, and Reliability Estimates for

Infantryman (11B) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Maintaining Supplies, Equipment, and Weapons	149	4.53	1.03	2.08 - 6.90	.63	172	4.55	0.78	2.37 - 6.29	.48
B. Assisting and Leading Others	148	4.06	1.06	1.65 - 6.15	.48	172	4.34	0.88	2.10 - 6.34	.49
C. Navigation	149	4.00	1.02	1.50 - 6.30	.47	172	4.22	0.98	2.12 - 6.52	.64
D. Use of Weapons and Other Equipment	149	4.73	0.97	1.67 - 6.96	.55	172	4.61	0.75	2.58 - 6.28	.55
E. Field Sanitation, Personal Hygiene, and Safety	149	4.77	1.04	1.18 - 7.00	.54	172	4.80	0.84	1.76 - 7.00	.43
F. Fighting Position	149	4.40	1.05	1.39 - 7.00	.49	172	4.33	0.86	1.99 - 6.03	.56
G. Avoiding Enemy Detection	149	4.28	1.01	1.58 - 6.30	.42	172	4.50	0.74	1.87 - 6.24	.30
H. Operating a Radio	149	4.64	1.05	2.22 - 6.96	.61	172	4.57	0.88	1.76 - 6.70	.60
I. Reconnaissance and Patrol	149	4.28	0.98	1.39 - 6.90	.55	172	4.43	0.88	1.90 - 6.44	.56
J. Guard and Security Duties	149	4.31	1.10	1.28 - 6.51	.53	172	4.60	0.81	2.32 - 6.48	.50
K. Courage and Proficiency in Battle	149	4.77	0.85	2.13 - 6.96	.33	172	4.63	0.85	2.26 - 6.47	.57
L. Prisoners of War	149	4.51	0.89	2.22 - 6.51	.29	172	4.32	0.81	2.24 - 6.03	.37
M. Overall Infantryman Performance	149	4.63	0.93	1.22 - 6.67	.53	172	4.76	0.79	2.12 - 6.70	.58
Mean Ratings	149	4.45	0.70	2.55 - 5.98	.78	172	4.5[illegible]	0.62	2.84 - 6.05	.81

Table 29

Supervisor and Peer Intercorrelations for

Infantryman (11B) MOS-Specific BARS

		Supervisors													Peers											
		A	B	C	D	E	F	G	H	I	J	K	L	M	A	B	C	D	E	F	G	H	I	J	K	L	M
	A. Maintain Weapons																										
S	B. Assist/Lead	65																									
U	C. Navigation	53	62																								
P	D. Use Weapons	59	54	56																							
E	E. Hygiene/Safety	49	40	33	45																						
R	F. Fighting Position	35	50	46	43	38																					
V	G. Avoid Detection	44	45	50	46	43	48																				
I	H. Operate Radio	43	40	44	37	23	30	36																			
S	I. Reconnaissance/Patrol	52	57	51	51	37	48	41	26																		
O	J. Guard/Security	46	49	35	32	37	41	42	19	39																	
R	K. Courage	48	50	53	55	39	59	49	35	56	31																
S	L. POW	32	27	30	47	26	31	43	30	35	19	44															
	M. Overall	68	64	55	56	46	54	44	51	58	51	56	39														
	A. Maintain Weapons	41	31	31	34	15	11	26	20	26	33	22	19	41													
	B. Assist/Lead	41	44	42	48	43	33	43	29	39	38	32	33	50	62												
	C. Navigation	43	49	47	43	33	37	52	43	37	36	41	33	44	61	58											
	D. Use Weapons	31	32	33	37	34	26	36	33	28	24	32	27	39	60	62	58										
P	E. Hygiene/Safety	32	25	31	34	39	25	32	23	31	24	30	24	33	46	47	39	44									
E	F. Fighting Position	37	39	35	39	42	46	37	19	37	21	43	34	43	38	50	57	48	41								
E	G. Avoid Detection	47	44	45	49	40	35	48	47	46	29	47	41	52	51	55	62	62	42	55							
R	H. Operate Radio	37	30	32	35	24	20	27	47	25	25	27	20	40	44	43	57	47	36	29	48						
S	I. Reconnaissance/Patrol	40	43	32	46	34	32	39	27	36	33	39	31	41	62	54	63	61	42	56	60	44					
	J. Guard/Security	28	28	34	37	34	29	30	24	35	30	32	12	33	45	52	43	51	54	40	36	36	54				
	K. Courage	34	29	39	37	29	30	38	22	35	19	39	36	37	49	49	52	56	38	52	50	47	62	48			
	L. POW	37	33	31	29	26	29	36	22	36	22	37	29	40	39	49	50	54	31	48	51	40	51	42	55		
	M. Overall	48	39	32	48	49	39	41	37	37	32	35	26	51	57	61	65	68	56	64	56	58	68	58	58	55	
		A	B	C	D	E	F	G	H	I	J	K	L	M	A	B	C	D	E	F	G	H	I	J	K	L	M

Armor Crewman - 19E

We tested 172 Armor Crewman enlistees during the Batch B field test sessions. Table 30 presents, for supervisor and peer ratings, means, standard deviations, range of ratings, and interrater reliability estimates.

We obtained complete supervisor rating data for 146 of these enlistees. Adjusted supervisor ratings range from 1.15 to 7.00. Mean adjusted ratings computed separately for each performance dimension range from 4.35 to 5.23 (standard deviations range from 0.72 to 1.15). The grand mean computed across all enlistees and performance dimensions, using the adjusted ratings, is 4.75 (SD= 0.58); for unadjusted ratings the grand mean is 4.89 (SD= 0.78). Interrater reliability estimates computed for each performance dimension range from .46 (E. Maintaining guns) to .73 (F. Engaging targets with tank guns) with a median value of .57.

We obtained complete peer rating data for 163 Armor Crewman enlistees. The adjusted values range from 1.45 to 7.00. Mean adjusted values computed for each performance dimension range from 4.38 to 5.01 with the standard deviations ranging from 0.71 to 0.98. The grand mean computed across all enlistees and performance dimensions, using the adjusted ratings, is 4.76 (SD= 0.56); the grand mean computed using unadjusted ratings is 4.75 (SD= 0.60). Interrater reliability estimates range from .29 (C. Stowing ammunition aboard tanks) to .65 (I. Overall performance) with a median value of .43.

Table 31 presents the intercorrelations for supervisor and peer ratings. For supervisor ratings alone, correlations for the first eight performance dimensions (excluding Overall performance) range from .09 to .47 with a mean value of .29 (SD= 0.11). According to supervisors, "Preparing tanks for field problems" (Dimension H), "Maintaining tank, tank systems, and associated equipment" (Dimension A), and "Engaging targets with tank guns" (Dimension F) correlate highest with "Overall performance."

Correlations between performance dimension ratings provided by peers (excluding Overall performance) range from .06 to .51, with a mean value of .35 (SD= 0.13). According to peers, "Preparing tanks for field problems" (Dimension H), "Engaging targets with tank guns" (Dimension F), and "Stowing ammunition aboard tanks" (Dimension C) correlate highest with "Overall performance."

Intercorrelations between peer and supervisor ratings computed for the first eight performance dimensions (excluding Overall performance) range from .02 to .42. Correlations appearing in the diagonal of this matrix range from .14 (C. Stowing ammunition aboard tanks) to .42 (F. Engage targets with tank guns) with a median value of .30.

Table 30

Means, Standard Deviations, Ranges, and Reliability Estimates for

Armor Crewman (19E) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Maintaining Tank, Tank Systems, and Associated Equipment	154	4.56	0.89	2.17 - 6.57	.55	163	4.57	0.78	1.85 - 6.46	.55
B. Driving and Recovering Tanks	153	4.67	1.07	1.15 - 6.57	.63	163	4.79	0.82	2.25 - 6.46	.42
C. Stowing Ammunition Aboard Tanks	154	5.08	0.72	2.57 - 6.76	.50	163	4.86	0.71	2.25 - 7.00	.29
D. Loading/Unloading Guns	154	5.23	0.90	2.01 - 7.00	.57	163	5.01	0.79	2.25 - 6.89	.37
E. Maintaining Guns	154	4.86	0.84	2.17 - 6.79	.46	163	4.83	0.79	2.83 - 6.49	.51
F. Engaging Targets with Tank Guns	146	4.35	1.15	1.57 - 6.60	.73	163	4.38	0.98	1.45 - 6.82	.51
G. Operating and Maintaining Communication Equipment	154	4.39	0.92	2.20 - 6.78	.57	163	4.51	0.80	1.81 - 6.81	.38
H. Preparing Tanks for Field Problems	154	4.79	0.94	1.22 - 6.78	.64	163	4.92	0.82	1.45 - 6.94	.43
I. Overall Armor Crewman Performance	153	4.85	0.82	2.17 - 7.00	.68	163	4.96	0.85	1.95 - 7.00	.65
Mean Ratings	154	4.75	0.58	2.56 - 6.11	.87	163	4.76	0.56	3.19 - 5.87	.76

Table 31

Supervisor and Peer Intercorrelations for

Armor Crewman (19E) MOS-Specific BARS

		Supervisors									Peers								
		A	B	C	D	E	F	G	H	I	A	B	C	D	E	F	G	H	I
S																			
U	A. Maintain Tank																		
P	B. Drive Tank	38																	
E	C. Stow Ammo	32	36																
R	D. Load/Unload	20	27	47															
V	E. Maintain Gun	30	14	24	40														
I	F. Engage Target	45	33	23	9	40													
S	G. Commo Equipment	15	10	11	27	36	24												
O	H. Prepare Tank	43	18	24	31	37	31	44											
R	I. Overall	53	42	27	31	44	50	42	54										
S																			
	A. Maintain Tank	39	14	17	13	23	34	8	23	22									
	B. Drive Tank	33	30	12	13	25	35	19	24	34	40								
P	C. Stow Ammo	25	7	14	23	27	28	17	29	23	50	35							
E	D. Load/Unload	24	14	17	22	28	26	2	17	32	34	33	51						
E	E. Maintain Gun	36	25	29	32	32	38	18	23	38	48	41	47	33					
R	F. Engage Target	34	18	24	30	38	42	29	23	44	42	33	38	36	47				
S	G. Commo Equipment	12	5	14	14	21	18	16	11	15	32	12	13	6	15	17			
	H. Prepare Tank	34	25	27	32	34	38	21	22	34	44	40	49	43	48	45	19		
	I. Overall	31	28	31	38	40	39	23	24	39	57	52	60	49	58	63	20	73	
		A	B	C	D	E	F	G	H	I	A	B	C	D	E	F	G	H	I

Radio Teletype Operator - 31C

In the field test sessions, we assessed the performance of 148 Radio Teletype Operator first-term enlistees. Means, standard deviations, range of ratings, and interrater reliability estimates are presented in Table 32.

According to the information in this table, we obtained complete supervisor rating data for 125 of those enlistees. Mean adjusted values computed across all enlistees for each performance dimension range from 4.26 to 4.93 (the standard deviation for these scores ranges from 1.01 to 1.16). The grand mean computed across all enlistees and performance dimensions, using adjusted ratings, is 4.68 (SD= 0.86); the grand mean for unadjusted ratings is 4.46 (SD= 0.93). Interrater reliability estimates range from .57 (C. Operating communications devices) to .70 (G. Overall performance) with a median value of .63.

From peers we obtained complete rating data for 120 Radio Teletype Operator enlistees. Mean adjusted values computed for each performance dimension range from 4.38 to 4.91 (standard deviations range from 0.85 to 1.03). The grand mean computed for adjusted ratings is 4.66 (SD= 0.69); the grand mean computed using unadjusted ratings is 4.88 (SD= 0.86). Interrater reliability estimates range from .52 (A. Inspecting and servicing equipment) to .69 (G. Overall performance) with a median value of .60.

Correlations computed between performance dimension ratings provided by supervisors and peers are shown in Table 33. For supervisors alone, these values range from .46 to .65 with a mean of .53 (SD= 0.05). (Values for the Overall rating are not included in the range or mean values above.) According to supervisors, "Installing and repairing equipment" (Dimension B), "Inspecting and servicing equipment" (Dimension A), and "Providing safe transportation" (Dimension F) are the dimensions most highly correlated with "Overall performance."

An examination of the peer data indicates that the correlations between the first seven performance dimensions (excluding Overall) range from .37 to .66 with a mean of .49 (SD= 0.09). According to peers, "Overall performance" correlates highest with performance in "Installing and repairing equipment" (Dimension B), "Operating communications devices" (Dimension C), and "Inspecting and servicing equipment" (Dimension A).

Intercorrelations computed between performance dimension ratings provided by peers and by supervisors (excluding Overall performance) range from .21 to .54. Correlations between supervisor and peer ratings on common performance dimensions range from .21 (C. Operating communications devices) to .63 (G. Overall performance) with a median value of .43.

Table 32

Means, Standard Deviations, Ranges, and Reliability Estimates for

Radio Teletype Operator (31C) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Inspecting and Servicing Equipment	125	4.26	1.08	1.00 - 6.43	.68	120	4.38	0.89	1.00 - 7.00	.52
B. Installing and Repairing Equipment	126	4.76	1.01	2.00 - 7.00	.63	121	4.64	0.85	2.78 - 7.00	.64
C. Operating Communications Devices	125	4.88	1.07	1.50 - 7.00	.57	121	4.84	0.86	2.00 - 6.43	.56
D. Preparing Reports	125	4.36	1.01	1.68 - 7.00	.58	121	4.41	1.03	1.76 - 7.00	.64
E. Maintaining Security	125	4.93	1.12	1.71 - 7.00	.67	121	4.91	0.88	2.78 - 7.00	.60
F. Providing Safe Transportation	125	4.83	1.11	1.00 - 7.00	.63	121	4.55	0.89	1.00 - 6.28	.53
G. Overall Radio Teletype Operator Performance	125	4.80	1.16	1.78 - 6.71	.70	121	4.80	1.02	1.26 - 6.76	.69
Mean Ratings	126	4.68	0.86	2.43 - 6.25	.80	121	4.66	0.69	2.78 - 6.54	.80

Table 33

Supervisor and Peer Intercorrelations for

Radio Teletype Operator - 31C MOS-Specific BARS

		Supervisors							Peers						
		A	B	C	D	E	F	G	A	B	C	D	E	F	G
S															
U															
P	A. Inspect Equipment														
E	B. Install Equipment	58													
R	C. Operate Commo	54	65												
V	D. Prepare Reports	54	46	48											
I	E. Maintain Security	52	52	52	46										
S	F. Provide Transportation	48	60	55	50	58									
O	G. Overall	67	69	58	59	58	63								
R															
S															
	A. Inspect Equipment	37	39	39	42	28	54	41							
P	B. Install Equipment	38	47	45	43	42	50	49	66						
E	C. Operate Commo	31	33	21	31	29	47	44	53	56					
E	D. Prepare Reports	39	38	41	43	38	48	37	44	62	39				
R	E. Maintain Security	32	26	34	29	41	46	27	51	48	37	37			
S	F. Provide Transportation	38	29	36	45	31	51	39	49	50	43	46	57		
	G. Overall	45	45	36	45	23	38	63	54	64	64	49	35	42	
		A	B	C	D	E	F	G	A	B	C	D	E	F	G

Light-Wheel Vehicle Mechanic - 63B

A total of 156 Light-Wheel Vehicle Mechanic enlistees were tested in the field test sessions. Data for these sessions are summarized in Table 34.

We obtained complete supervisor rating data for 137 of these enlistees. Mean adjusted scores computed across all enlistees for each performance dimension range from 3.96 to 4.92 (standard deviations for these ratings range from 1.03 to 1.23). The grand mean computed across all enlistees and performance dimensions, for adjusted ratings, is 4.48 (SD= 0.87); the grand mean computed for unadjusted ratings is 4.34 (SD= 0.98). Estimates of interrater reliability range from .43 (C. Performing routine maintenance) to .67 (L. Overall performance) with a median value of .62.

From peers we obtained complete data for a total of 127 Light-Wheel Vehicle Mechanic enlistees. Mean adjusted values computed for each performance dimension range from 4.11 to 4.92 (standard deviations range from 0.94 to 1.12). The grand mean computed for adjusted ratings is 4.47 (SD= 0.73); using unadjusted ratings the grand mean is 4.64 (SD= 0.81). Interrater reliability estimates range from .35 (K. Recovery) to .70 (C. Performing routine maintenance) with a median value of .59.

Table 35 contains the intercorrelations computed between performance dimension ratings for supervisors and peers. For supervisors correlations among the first 11 performance dimensions (excluding Overall performance) range from .31 to .77 with a mean of .53 (SD= .10). Performance dimension ratings yielding the highest correlations with "Overall performance" for the supervisor group include "Troubleshooting" (Dimension B), "Performing routine maintenance" (Dimension C), "Inspecting, testing, and detecting problems with equipment" (Dimension A), and "Repair" (Dimension D).

Correlations between performance dimension ratings provided by peers (excluding Overall performance) range from .08 to .69 with a mean value of .43 (SD= 0.13). Peers agree with supervisors that "Repair" (Dimension D), "Troubleshooting" (Dimension B), and "Inspecting, testing, and detecting problems with equipment" (Dimension A) correlate highest with "Overall performance".

Intercorrelations between performance dimension ratings provided by supervisors and peers (excluding Overall) range from .06 to .57. Correlations in the diagonal of supervisor-peer matrix range from .26 (K. Recovery) to .62 (L. Overall performance) with a median value of .45.

Table 34

Means, Standard Deviations, Ranges, and Reliability Estimates for

Light-Wheel Vehicle Mechanic (63B) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Inspecting, Testing, and Detecting Problems with Equipment	141	4.31	1.14	1.00 - 6.96	.64	128	4.39	1.02	1.67 - 7.00	.68
B. Troubleshooting	141	4.19	1.16	1.00 - 7.00	.63	128	4.26	1.05	1.00 - 6.83	.63
C. Performing Routine Maintenance	142	4.86	1.03	1.43 - 7.00	.43	128	4.85	1.04	1.43 - 7.00	.70
D. Repair	142	4.71	1.12	1.00 - 7.00	.57	128	4.69	1.04	1.99 - 7.00	.66
E. Using Tools and Test Equipment	142	4.45	1.21	1.00 - 7.00	.59	128	4.40	1.01	1.00 - 6.43	.40
F. Using Technical Documentation	142	4.58	1.23	1.00 - 7.00	.66	128	4.37	0.94	1.33 - 7.00	.42
G. Vehicle and Equipment Operation	141	4.92	1.12	1.43 - 7.00	.50	128	4.92	1.00	1.00 - 7.00	.63
H. Safety Mindedness	142	4.60	1.08	1.98 - 7.00	.64	128	4.32	0.95	2.25 - 6.43	.49
I. Administrative Duties	143	4.03	1.21	1.08 - 7.00	.57	128	4.25	1.12	1.00 - 7.00	.64
J. Planning/Organizing Jobs	143	3.96	1.17	1.00 - 7.00	.61	128	4.11	1.05	1.33 - 6.28	.42
K. Recovery	137	4.51	1.17	1.00 - 7.00	.62	127	4.34	0.98	1.17 - 6.42	.35
L. Overall Light-Wheel Vehicle Mechanic Performance	142	4.69	1.14	1.00 - 7.00	.67	128	4.76	1.04	1.08 - 7.00	.54
Mean Ratings	143	4.48	0.87	1.33 - 6.67	.85	128	4.47	0.73	2.42 - 6.20	.86

Table 35

Supervisor and Peer Intercorrelations for

Light-Wheel Vehicle Mechanic - 63B MOS-Specific BARS'

		Supervisors												Peers											
		A	B	C	D	E	F	G	H	I	J	K	L	A	B	C	D	E	F	G	H	I	J	K	L
S	A. Inspect Equipment																								
U	B. Troubleshoot	77																							
P	C. Maintenance	72	69																						
E	D. Repair	66	69	67																					
R	E. Use Tools	46	63	49	50																				
V	F. Use Technical Docments	55	60	58	54	62																			
I	G. Vehicle Operation	57	50	63	47	44	47																		
S	H. Safety	44	50	49	50	55	58	53																	
O	I. Administrative Duties	51	53	46	44	55	59	50	49																
R	J. Plan/Organize	53	68	56	61	64	70	45	52	64															
S	K. Recovery	43	48	43	46	31	33	43	35	36	40														
	L. Overall	68	73	70	68	61	55	50	49	48	65	52													
	A. Inspect Equipment	51	44	48	41	46	47	30	34	43	48	20	56												
	B. Troubleshoot	45	45	50	47	43	50	43	46	32	55	33	56	65											
	C. Maintenance	45	46	44	46	38	44	34	45	43	53	37	54	67	55										
P	D. Repair	45	47	53	49	46	49	45	45	37	51	44	55	59	69	65									
E	E. Use Tools	33	37	33	37	43	41	41	55	30	38	27	39	31	50	35	53								
E	F. Use Technical Documents	35	40	38	29	43	48	29	32	33	45	7	42	54	49	34	50	26							
R	G. Vehicle Operation	41	47	39	37	45	49	28	38	41	46	21	42	55	53	59	52	45	38						
S	H. Safety	31	35	36	31	38	42	28	30	35	33	6	34	27	32	23	38	38	31	39					
	I. Administrative Duties	45	50	57	43	50	57	52	50	45	55	29	52	46	51	46	48	39	52	48	46				
	J. Plan/Organize	29	30	37	23	43	49	32	38	41	43	16	37	54	56	51	51	40	35	47	41	60			
	K. Recovery	40	30	38	21	[illegible]	28	23	10	25	31	26	40	39	36	29	31	15	23	32	8	27	28		
	L. Overall	54	54	55	58	48	56	49	46	40	52	34	62	65	66	59	68	43	56	54	31	51	47	34	
		A	B	C	D	E	F	G	H	I	J	K	L	A	B	C	D	E	F	G	H	I	J	K	L

Medical Specialist - 91A

A total of 167 Medical Specialist enlistees were included in the field test sessions. Data for this MOS are summarized in Table 36.

As Table 36 indicates, we obtained complete supervisor rating data from 138 of these enlistees. Adjusted mean scores computed across all enlistees for each performance dimension range from 4.39 to 5.17 (standard deviations for these values range from 0.97 to 1.24). The grand mean computed across all enlistees and performance dimensions, using adjusted ratings, is 4.71 (SD= 0.79); for unadjusted ratings the grand mean is 4.71 (SD= 0.83). Interrater reliability estimates range from .45 (G. Providing routine and ongoing patient care) to .75 (C. Keeping medical records) with a median value of .66.

We obtained complete peer rating data for 148 Medical Specialists. Adjusted mean values computed across all enlistees for each performance dimension range from 4.45 to 4.93 (standard deviations range from 0.84 to 1.03). The grand mean computed using the adjusted ratings is 4.71 (SD= 0.72); across the unadjusted ratings the grand mean is 4.72 (SD= 0.76). Interrater reliability estimates computed for peers range from .44 (F. Preparing and inspecting field site or clinic facilities) to .68 (I. Providing health care and health maintenance instructions to Army personnel) with a median value of .62.

Correlations between performance dimension ratings provided by supervisors and peers are provided in Table 37. For supervisors alone, values for the first nine dimensions (excluding Overall performance) range from .25 to .57 with a mean value of .45 (SD= 0.08). According to supervisors, "Responding to emergency situations" (Dimension H), "Keeping medical records" (Dimension C), and "Maintaining accountability of medical supplies and equipment" (Dimension B) correlate highest with "Overall performance."

Focusing on peer rating data, correlations between ratings on the first nine performance dimensions (excluding Overall performance) range from .33 to .70 with a mean value of .53 (SD= 0.09). According to peers, "Responding to emergency situations" (Dimension H), "Dispensing medication" (Dimension E), and "Providing routine and ongoing patient care" (Dimension G) correlate highest with "Overall performance."

Intercorrelations among supervisor and peer ratings across all performance dimensions, excluding "Overall performance", range from .18 to .57. Correlations computed between supervisor and peer ratings on common performance dimensions range from .29 (F. Preparing and inspecting field site or clinic facilities) to .59 (J. Overall performance) with a median value of .43.

Table 36

Means, Standard Deviations, Ranges, and Reliability Estimates for

Medical Specialist (91A) MOS-Specific BARS - Supervisors and Peers

	Supervisors					Peers				
Scale	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Maintaining and Operating Army Medical Vehicles and Equipment	154	4.40	1.24	1.00 - 6.72	.70	148	4.61	0.95	1.40 - 7.00	.45
B. Maintaining Accountability of Medical Supplies and Equipment	156	4.48	1.22	1.63 - 7.00	.62	157	4.54	0.91	1.45 - 7.00	.65
C. Keeping Medical Records	144	4.53	1.11	1.13 - 6.69	.75	158	4.62	1.03	1.00 - 6.80	.66
D. Arranging for Transportation and/or Transporting Injured Personnel	148	4.89	1.00	1.72 - 7.00	.48	155	4.73	0.94	1.70 - 7.00	.60
E. Dispensing Medications	138	4.79	0.97	2.17 - 7.00	.61	155	4.86	0.89	1.40 - 6.90	.55
F. Preparing and Inspecting Field Site or Clinic Facilities	146	4.39	1.11	1.77 - 7.00	.68	150	4.45	0.84	1.64 - 6.80	.44
G. Providing Routine and Ongoing Patient Care	152	5.00	1.05	2.22 - 7.00	.45	158	4.89	0.94	1.45 - 7.00	.64
H. Responding to Emergency Situations	154	5.17	1.14	1.22 - 7.00	.72	158	4.93	0.94	1.45 - 7.00	.54
I. Providing Health Care and Health Maintenance Instructions to Army Personnel	155	4.54	1.12	1.13 - 7.00	.63	158	4.45	0.98	1.40 - 6.70	.68
J. Overall Medical Specialist Performance	156	4.89	1.04	1.22 - 7.00	.69	158	4.88	0.91	1.45 - 7.00	.67
Mean Ratings	156	4.71	0.79	1.82 - 6.50	.85	158	4.71	0.72	1.70 - 6.49	.84

Table 37

Supervisor and Peer Intercorrelations for

Medical Specialist - 91A MOS-Specific BARS

		Supervisors										Peers									
		A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
S	A. Maintain Vehicles/Equipment																				
U	B. Account for Supplies/Equipment	53																			
P	C. Keep Records	47	57																		
E	D. Transportation	47	40	39																	
R	E. Dispense Medication	46	35	52	45																
V	F. Prepare Field Sites	53	50	36	56	46															
I	G. Patient Care	38	37	51	25	43	26														
S	H. Emergencies	34	38	46	53	56	34	52													
O	I. Health Instruction	43	49	57	40	56	49	45	43												
R	J. Overall	58	64	65	58	58	54	54	70	60											
S																					
	A. Maintain Vehicles/Equipment	45	21	30	33	34	29	22	43	36	44										
	B. Account for Supplies/Equipment	40	42	53	28	37	29	39	43	52	52	65									
	C. Keep Records	37	38	56	18	41	19	44	34	47	46	43	62								
P	D. Transportation	36	39	37	34	39	32	24	40	40	41	53	58	33							
E	E. Dispense Medication	31	37	40	26	38	19	32	45	47	47	53	56	55	43						
E	F. Prepare Field Sites	36	40	31	25	21	29	28	35	30	42	49	46	35	48	53					
R	G. Patient Care	40	40	54	22	40	25	41	43	50	50	50	60	61	47	68	45				
S	H. Emergencies	39	34	48	34	47	21	40	57	48	53	53	59	52	54	65	43	70			
	I. Health Instruction	42	33	38	26	40	28	42	47	44	51	55	44	49	47	63	57	64	65		
	J. Overall	44	46	53	34	42	29	46	43	55	59	55	64	57	52	71	54	70	72	67	
		A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J

Discussion and Conclusions

Analyses of the field test data indicate that peers and supervisors provided useful information about MOS-specific job performance, with each rater group providing unique information about MOS-specific job requirements.

Supervisor and peer ratings yielded similar levels of reliability estimates. Across all MOS, median reliability estimates for supervisor ratings range from .53 for Infantryman (11B) to .66 for Medical Specialist (91A) with a median value of .57. For peer ratings, median values range from .43 for Armor Crewman (19E) to .65 for Military Police (95B) with a median value of .55. The median values indicate that for single item scales, interrater reliability estimates are at acceptable levels. Median values for the two rater type groups suggest that supervisors are probably more reliable than peers. Recall that assumptions for computing interrater reliability estimates differed for supervisors and peers; we assumed three or four peer raters for each ratee and two supervisor raters for each ratee. Reported reliability estimates were adjusted for the number of raters for each ratee. Given equal numbers of supervisor and peer raters for each ratee, these data indicate that the supervisor ratings would be somewhat more reliable than the peer ratings.

Supervisors and peers provided similar information about the mean level of performance. Across the nine MOS, peers provided slightly higher grand mean values than supervisors in two MOS, Administrative Specialist (71L) and Infantryman (11B). Supervisors provided slightly higher grand mean values than peers in two MOS, Motor Transport Operator (64C) and Military Police (95B). Mean ratings for the two groups were nearly identical for the remaining MOS, Cannon Crewman (13B), Armor Crewman (19E), Radio Teletype Operator (31C), Light-Wheel Vehicle Mechanic (63B), and Medical Specialist (91A).

Average intercorrelations among performance dimension ratings for supervisors and peers are similar. For supervisor ratings, the mean correlation for the nine MOS ranges from .29 for Armor Crewman (19E) to .53 for Radio Teletype Operator (31C) and Light-Wheel Vehicle Mechanic (63B). For peer ratings, the mean correlation across the nine MOS ranges from .35 for Armor Crewman (19E) to .58 for Military Police (95B). The greatest difference between mean correlations for supervisors and peers occurs for Military Police (95B) with the mean value for supervisors at .39 and mean value for peers at .58.

For each MOS, we identified three performance dimensions ratings that in the judgment of supervisors and peers correlated highest with the "Overall performance" rating. This information suggests how the two rater groups differ with respect to perceptions about requirements that lead to success on the job. Across the nine MOS, correlations between performance dimension ratings and the "Overall performance" rating indicate that supervisors and peers agree only moderately on the requirements that lead to success on the job.

For four MOS, Administrative Specialist (71L), Armor Crewman (19E), Radio Teletype Operator (31C), and Light-Wheel Vehicle Mechanic (63B), peers and supervisors agreed on two of the three performance dimensions contributing most to overall performance. For three MOS, Cannon Crewman (13B), Military Police (95B), and Medical Specialist (91A), supervisors and peers agreed on one of three performance dimensions. For two MOS, Motor Transport Operator (64C) and Infantryman (11B), there was no agreement among supervisors and peers concerning the performance dimensions that correlate highest with "Overall Performance."

Finally, correlations computed between supervisor and peer ratings on common performance dimensions reveal a moderate amount of agreement between the two rater groups. Median correlations computed for each MOS range from .30 for Armor Crewman (19E) to .46 for Motor Transport Operators (64C).

In sum, supervisors and peers provided performance ratings that were similar in reliability, mean performance level, and average intercorrelation between performance dimensions. Supervisors and peers, however, appeared to differ somewhat in their perceptions of requirements that lead to overall success on the job.

CHAPTER 3: PREPARATION OF THE MOS-SPECIFIC BARS FOR ADMINISTRATION IN THE CONCURRENT VALIDITY STUDY

Prior to administering the MOS-specific rating scales in the Concurrent Validity study, scale developers reviewed results from the field test data analyses. Further, the MOS-specific rating scales were submitted to a Proponent review to verify that critical first-term job requirements were represented in the performance scales. In this chapter we describe the procedures for modifying the MOS-specific behaviorally anchored rating scales, using results from the field test as well as input supplied by the Proponent review committee.

Evaluation of Field Test Results

Reliability

In Chapter 2, we summarized the reliability estimates computed for supervisor and peer ratings obtained from the field test sessions. Although we concluded that, on the average, single-scale reliability estimates were acceptable for each rater group, we were concerned that within a particular MOS there might be one or two performance dimensions on which supervisors and peers alike experienced difficulty in evaluating enlistees. Consistently low reliability estimates observed for both rater groups on a particular performance dimension might suggest that the dimension definition and anchors were unclear or that the dimension did not reflect a critical component of the job.

For each MOS, we compared the reliability estimates computed for performance dimension ratings provided by supervisors with estimates for ratings provided by peers to identify possible problem dimensions. Table 38 provides a summary of the median reliability estimates as well as the range of reliabilities for each MOS.

For most MOS, there appears to be no consistent pattern when reliability estimates computed for supervisor ratings are compared with those computed for peer ratings. In only one MOS, Military Police (95B), the pattern of reliability estimates for supervisor ratings and peer ratings corresponded quite highly. Within that MOS one performance dimension, "Providing security" (Dimension B), appeared to present problems for both rater groups. The interrater reliability estimate computed separately for supervisors and peers is the same for both groups, .39. Therefore, we reviewed this particular performance dimension to clarify the definition as well as the behavioral anchors.

For the remaining MOS-specific rating scales, we identified performance dimensions with low reliability estimates computed for peer or supervisor ratings. We then reviewed rating scale definitions and anchors developed for these dimensions to uncover potential problems.

Table 38

MOS-Specific BARS: Summary of Reliability Estimates for Supervisor and Peer Ratings

	Supervisors			Peers		
MOS	Median	Range		Median	Range	
13B Cannon Crewman	.54	.33	J. Position Improvement	.54	.40	H. Receiving and Relaying Communications
		.70	K. Overall Performance		.66	G. Load/Unload Howitzers
64C Motor Transport Operator	.57	.47	F. Parking and Securing Vahicles	.54	.32	G. Performing Administrative Duties
		.66	I. Safety Mindedness and E. Loading Cargo/Transporting Personnel		.68	D. Using Maps and Following Proper Routes
71L Administrative Specialist		Not Calculated - Insufficient Number of Pairs		.46	.37	H. Providing Customer Service
					.55	G. Safeguarding and Monitoring Security of Classified Documents and I. Overall Performance
95B Military Police	.55	.39	B. Providing Security	.65	.39	B. Providing Security
		.74	H. Overall Performance		.71	H. Overall Performance
11B Infantryman	.53	.29	L. Prisoners of War	.55	.30	G. Avoiding Enemy Detection
		.63	A. Maintaining Supplies, Equipment and Weapons		.64	C. Navigation
19E Armor Crewman	.57	.46	E. Maintaining Guns	.43	.29	C. Stowing Ammunition Aboard Tanks
		.73	F. Engaging Targets with Tank Guns		.65	I. Overall Performance
31C Radio Teletype Operator	.63	.57	C. Operating Communication Devices	.60	.52	A. Inspecting and Servicing Equipment
		.70	G. Overall Performance		.69	G. Overall Performance
63B Light-Wheel Vehicle Mechanic	.62	.43	C. Performing Routine Maintenance	.59	.35	K. Recovery
		.67	L. Overall Performance		.70	C. Performing Routine Maintenance
91A Medical Specialist	.66	.45	G. Providing Routine and Ongoing Patient Care	.62	.44	F. Preparing and Inspecting Field Site or Clinic Facilities
		.75	C. Keeping Medical Records		.68	I. Providing Health Care Instruction to Army Personnel

Leniency and Severity

As reported in Chapter 2, we computed grand mean values separately for peer ratings and supervisor ratings; for the two rater type groups these mean values are very similar. We used these values to assess leniency and severity effects. High mean values indicate that raters may have been too lenient or "easy" in assigning ratings, whereas very low mean values indicate that raters may have been too severe or strict in assigning ratings.

Recall that the grand mean values tabulated in Chapter 2 were computed using adjusted ratings. Grand means computed using the raw rating data provide a more appropriate statistic for evaluating ratings for leniency or severity effects. Table 39 contains the grand mean values reported by MOS and by rater type. Grand mean values computed using both the unadjusted and adjusted ratings have been included for comparison purposes.

Grand mean values computed using adjusted scores correspond very highly with those values computed using unadjusted scores. For supervisors the grand mean values, using unadjusted ratings, range from 4.34 to 4.92; for adjusted ratings these values range from 4.48 to 5.07. For peers the grand mean values for unadjusted ratings range from 4.43 to 4.89; for adjusted ratings the values range from 4.43 to 4.85.[2]

Since the scale used for making these ratings ranges from 1 (low or ineffective performance) to 7 (high or effective performance), one might argue that ratings which reflect no leniency or severity effects should be near 4.00. According to the results from the field test, grand means computed across individual performance dimensions separately for each MOS and rater type are all above 4.00. One might conclude, then, that these data demonstrate leniency effects.

Cascio and Valenzi (1978), however, argue that ratings which appear lenient might, in fact, accurately reflect incumbents' job performance, because prior selection has weeded out potentially poor performers. Supervisor and peer ratings obtained in the field test sessions do not appear overly lenient and may, in fact, reflect job performance levels we would expect, given that poorer performers have been identified and screened out through the selection and classification process as well as through Basic Training and Advanced Individual Training.

Proponent Review Procedures and Results

Following the Batch B field test administration, each of the nine MOS-specific behaviorally anchored rating scales was submitted to a Proponent committee for review. Proponent committee members, who were primarily technical school subject matter experts from each MOS, studied the scales and made suggestions for scale modifications.

---

[2]Unadjusted and unscreened rating data provided by supervisors and peers are summarized in Section 5 of the nine MOS appendices.

Table 39

Summary of Grand Mean Values for Unadjusted and Adjusted Ratings by MOS[a]

		Supervisors		Peers	
MOS		Unadjusted	Adjusted	Unadjusted	Adjusted
13B Cannon Crewman	Mean	4.89 (1.13)	4.89 (0.81)	4.89 (0.84)	4.85 (0.71)
	Median	4.90	4.92	4.97	4.92
64C Motor Transport Operator	Mean	4.92 (1.02)	5.07 (0.73)	4.66 (0.83)	4.74 (0.66)
	Median	5.00	4.85	4.78	4.88
71L Administrative Specialist	Mean	4.56 (1.13)	4.52 (0.94)	4.75 (0.81)	4.72 (0.64)
	Median	4.57	4.52	4.79	4.73
95B Military Police	Mean	4.59 (0.75)	4.47 (0.63)	4.43 (0.66)	4.43 (0.60)
	Median	4.59	4.58	4.41	4.47
11B Infantryman	Mean	4.39 (0.91)	4.45 (0.70)	4.56 (0.70)	4.51 (0.62)
	Median	4.44	4.51	4.60	4.55
19E Armor Crewman	Mean	4.89 (0.78)	4.75 (0.58)	4.75 (0.60)	4.76 (0.56)
	Median	4.91	4.79	4.84	4.83
31C Radio Teletype Operator	Mean	4.46 (0.93)	4.68 (0.86)	4.88 (0.86)	4.66 (0.69)
	Median	4.57	4.80	4.87	4.64
63B Light-Wheel Vehicle Mechanic	Mean	4.34 (0.98)	4.48 (0.87)	4.64 (0.81)	4.47 (0.73)
	Median	4.41	4.59	4.54	4.38
91A Medical Specialist	Mean	4.71 (0.83)	4.71 (0.79)	4.72 (0.76)	4.71 (0.72)
	Median	4.70	4.67	4.70	4.68

[a]Standard deviations are shown in parentheses.

For most MOS, suggestions made by committee members included minor wording changes. For example, committee members noted a problem with one of the anchors in one Administrative Specialist (71L) performance dimension, "Keeping records." Specifically, the committee recommended deleting one anchor from this dimension because it described job duties typically required of second-term personnel only (i.e., handle suspense dates). Therefore, we omitted this anchor from that performance dimension.

For another MOS, Radio Teletype Operators (31C), the Proponent review committee noted that the job title had been changed. Therefore, we made the necessary changes on all Concurrent Validity study rating forms. The current MOS-Specific rating form for this MOS now reads "Single Channel Radio Operator--31C."

For one MOS, Military Police (95B), the committee asked for more extensive changes. Committee members noted that because critical incident workshops were conducted only in CONUS locations, a few requirements of the Military Police job were missing. Incumbents in this MOS serving in OCONUS locations are required to provided combat and combat support functions. Thus, four performance dimensions describing these requirements were added to the Military Police MOS-specific rating scales: (1) "Navigation" (Dimension H); (2) "Avoiding enemy detection" (Dimension I); (3) "Use of weapons and other equipment" (Dimension J); and (4) "Courage and proficiency in battle" (Dimension K). Definitions and behavioral anchors for these scales had been developed for the Infantryman (11B) performance dimensions rating scales. Proponent committee members reviewed these definitions and anchors and authorized including the same information in the Military Police performance rating scales.

Project-Wide Review Committee

Following the Batch B field test sessions, Project A staff members reviewed the final set of rating scales. This group, the Criterion Measurement Task Force, was composed of project personnel responsible for developing task-oriented and behavior-oriented criterion measures. Further, most members had participated in administering criterion measures during the Batch A and Batch B field tests.

Task Force participants reported that some of the rating scales, the behaviorally anchored scales in particular, required considerable reading time. Consequently, they believed that many raters were not reading the scales thoroughly before making their ratings. This group recommended that we pare down the length of the behavioral anchors to help ensure that all raters would review the anchors thoroughly before using them to evaluate incumbents.

Therefore, PDRI staff responsible for developing the nine MOS-specific ratings scales modified the performance dimension definitions and scale anchors. Their goal was to retain the specific job requirements and depiction of ineffective, adequate, or effective performance in each anchor while eliminating unnecessary information or lengthy descriptions. Figure 5 contains an example of the anchors for one performance dimension included

Before - BATCH A Field Test Administration

A. TRAFFIC CONTROL AND ENFORCEMENT

Controlling traffic and enforcing traffic laws and parking rules.

1 2	3 4 5	6 7
Below Standard	Adequate/Mid-Range	Superior
. Often uses hand/arm signals that are difficult to understand, at times resulting in unnecessary accidents; often fails to wear reflectorized gear; overlooks hazardous traffic *conditions; sleeps on duty; pays* excessive attention to things unrelated to the job. . May display excess leniency or harshness when citing offenders, allowing their military rank, race, and/or sex to influence his/her actions; makes many errors when filling out citations.	. Usually does a reasonable job when directing traffic by using adequate hand/arm signals and/or wearing reflectorized gear. . Makes few errors when filling out citations; usually does not allow an offender's race, sex, and/or military rank to interfere with good judgment.	. Consistently uses appropriate hand/arm signals; always wears reflectorized gear; generally monitors traffic from plain-view vantage points; consistently refrains from behaviors such as reading and prolonged conversation on non-job related topics. . Always uses emergency equipment (e.g., flares, barricades) to highlight unsafe conditions and ensures that hazards are removed or otherwise taken care of.

After - Concurrent Validity Study

A. TRAFFIC CONTROL AND ENFORCEMENT

How effective is each soldier in controlling traffic and enforcing traffic laws and parking rules?

① ②	③ ④ ⑤	⑥ ⑦
Often uses hand/arm signals that are difficult to understand or fails to wear reflectorized gear; overlooks many hazardous traffic conditions and violations. May allow the military rank, race, or sex of traffic offenders to influence his/her enforcement of traffic laws; makes many errors when filling out citations.	Generally uses adequate hand/arm signals and wears reflectorized gear *when directing traffic; usually pays* attention to traffic conditions and enforces traffic laws. Usually does not allow a traffic offender's military rank, race, or sex to influence his/her enforcement when filling out citations.	Always uses adequate hand/arm signals and wears reflectorized gear when directing traffic; monitors traffic carefully from plain-view vantage points and enforces all traffic laws. Never allows a traffic offender's military rank, race, or sex to influence his/her enforcement of traffic laws; rarely or never makes errors when filling out citations.

Figure 5. Example Performance Rating Scale from Military Police (95B) MOS-Specific BARS, Before and After Modifications

in the Military Police (95B) rating scales as they appeared for the Batch B administration and as they appear for the Concurrent Validity study.

The rating scales to be administered in the Concurrent Validity study have been included in Section 6 of the nine MOS appendices to this report.

Concurrent Validity Study Plans

Administration

Throughout field test data collection efforts, PDRI staff members conducting rating sessions identified problems with particular rating instruments and ways to improve the rating sessions. This information was summarized in memos to the various task leaders.

In sum, rating session administrators reported few or no problems with the MOS-specific rating scales. The only complaint with these particular scales was that they did not offer a "Cannot Rate" option for raters who feel unable to evaluate an incumbent on a particular performance dimension. We decided that for the Concurrent Validity study, we would not include a "Cannot Rate" option. Instead, rating session administrators would be instructed to encourage raters to evaluate ratees on ALL performance dimensions. Raters who simply could not evaluate a ratee on a particular dimension would be asked to leave that scale blank. (For a complete description of guidelines provided to rating session administrators for the Concurrent Validity study, see Pulakos & Borman, 1986.)

Data Analysis

Data analyses for Batch A and Batch B field test data have been described in Chapter 2 of this report. Briefly, this process entailed computing adjusted rating scores for raters using information from supervisors and peers combined; following the adjustment procedures, we analyzed supervisor and peer rating data separately.

Data collected in the Concurrent Validity study with a larger sample size for each MOS will permit additional analyses that were not performed on the field test data. These include the following:

- Compare adjusted scores with unadjusted scores to determine whether one procedure is better than the other in terms of reliability, halo, and rating score distributions.
- Factor analyze intercorrelations computed between performance dimension ratings provided by supervisors. Compare the resulting factors with factors obtained from the peer rating data.
- Determine whether or how to best combine the information supplied by supervisors and peers.

- Examine correlations between ratings obtained on MOS-specific rating scales and criterion data obtained on other measures (e.g., hands-on tests, job knowledge tests, training knowledge tests). This information would provide a clearer understanding of the job performance components that we are capturing in the MOS-specific BARS. Further, these data would be useful in developing criterion composite measures.

Summary

In this chapter, we described the information used to modify the MOS-specific behaviorally anchored rating scales developed for nine MOS, prior to their use in the Concurrent Validity study. Briefly, we relied on information obtained from field test administrations, recommendations provided by subject matter experts, and suggestions offered by project staff.

In general, very few content changes were made on the rating scales, with the exception of additional scales developed for Military Police (95B) to reflect overseas requirements. Across all MOS-specific rating scales, however, we pruned the behavioral anchors to reduce reading requirements while maintaining the flavor and standards depicted in each anchor.

REFERENCES

Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 60, 412-421.

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1987). Development of a model of soldier effectiveness (ARI Technical Report 741).

Borman, W. C., & Rose, S. R. (1986). Chapter 2: Development of the Army-wide rating scales and task dimensions. In E. D. Pulakos & W. C. Borman (Eds.), Development and field test of Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716). (AD B112 857)

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). Development and field test of Project A task-based MOS-specific criterion measures (ARI Technical Report 717). (AD A182 645)

Campbell, J. P., Dunnette, M. D., Arvey, R., & Hellervik, L. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

Cascio, W. F., & Valenzi, E. R. (1978). Relations among criteria of police performance. Journal of Applied Psychology, 63, 22-28.

Davis, R. H., Davis, G., Joyner, J., & de Vera, M. V. (1985). Development and field test of job-relevant knowledge tests for selected MOS (ARI Technical Report 776).

Eaton, N. K., & Goer, M. H. (Eds.). (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Technical appendix to the annual report (ARI Research Note 83-37). (AD A137 117)

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (Eds.). (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year (ARI Technical Report 660). (AD A178 944)

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report (ARI Research Report 1347). (AD A141 807)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Project A - Research plan (ARI Research Report 1332). (AD A129 728)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report synopsis, 1984 fiscal year (ARI Research Report 1393). (AD A173 824)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Appendices to annual report, 1984 fiscal year (ARI Technical Report 660). (AD A178 944)

Olson, D. M., & Borman, W. C. (1987). Development and field tests of the Army Work Environment Questionnaire (ARI Technical Report 737). (AD A182 078)

Peterson, N. G. (Ed.). (1987). Development and field test of the trial battery for Project A (ARI Technical Report 739). (AD A184 575)

Pulakos, E. D. (1986). Chapter 6: Batch B rater training experiment: The effects of practice on making ratings. In E. D. Pulakos & W. C. Borman (Eds.), Development and field test of Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716). (AD B112 857)

Pulakos, E. D., & Borman, W. C. (1986). Chapter 5: Rater orientation and training. In E. D. Pulakos & W. C. Borman (Eds.), Development and field test report for the Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716). (AD B112 857)

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.